Causality inference between time series data and its applications

Siyuan Chen

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020

Abstract

Causality Inference Between Time Series Data and Its Applications

Siyuan Chen

Ever since Granger[2] first proposed the idea of quantitatively testing the causal relationship between data streams, the endeavor of accurately inferring the causality in data and using that information to predict the future has not stopped. Artificial Intelligence (AI), by utilizing massive amount of data, helps to solve complex problems, whether they include the diagnosis and detection of disease through medical imaging, email spam detection, or self-driving vehicles. Perhaps, this thesis will be trivial in ten years from now. AI has pushed human kind to reach the next technological level in technology. Nowadays, among most machine leaning inquiries, statistical relationships are determined using correlation measures. By feeding data into machine learning algorithms, computers update the algorithm's parameters iteratively by extracting and mapping features to learning targets until the correlation increases to a significant level to cease the training process. However, with the increasing developments of powerful AI, there is really a shortage in exploring causality in data. It is almost self-evident that "correlation is not causality"[1]. Sometimes, the strong correlation established between variables through machine learning can be absurd and meaningless. Providing insight into causality information through data, which most of the machine learning methods fall short to do, is of paramount importance. The subsequent chapters detail the four endeavors of studying causality in financial markets, earthquakes, animal/human brain signals, the predictivity of data sets. In Chapter 2, we further developed the concept of causality networks[11] into higher order causality network. We applied

these to financial data, and tested their validity and ability to capture the system's causal relationship. In next Chapter 3, We examined another type of time series-earthquakes. Violent seismic activities decimate people lives, and destroy entire cities and areas. This begs us to understand how earthquakes work and help us make reliably and evacuation-actionable predictions. The causal relationships of seismic activities in different areas are studied and established. Biological data, specifically brain signals, are time series data and their causal pattern are explored and studied. Different human and mice brain signals are analyzed and clustered in Chapter 4 using their unique causal pattern to understand different brain cell activity. Finally, we realized that the causal pattern in the time series can be used to compress data. A causal compression ratio is invented and used as the data stream's predictivity index. We describe this in Chapter 5.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I thought this would be the easiest chapter to write, but it turned out to be the hardest.

First and foremost, I want to thank my advisor, Professor **Hod Lipson**. I never believed I was a good enough student to be admitted, but I feel so lucky that he took me under his supervision and that I had the chance to join the **creative machines lab**. He showed me his intelligence, prudence, and patience. He taught me what good research work and what a good researcher should look like. He always said: Start with a good and novice idea, validate it carefully, and be able to communicate your resultseven to a cable driver. Despite his busy schedule, he always had time and patience for me. I began by doing evolutionary robotic work, then learned about precision agriculture, and finally, landed on solving the mystery of the causal relationship of the time series data stream. Professor Hod Lipson pushed me to be creative and autonomous. He has taught me many invaluable qualities that I will carry with me throughout my life.

I want to also thank Melbourne E. Francis and Aixa L. Rosado for their administrative work. I wouldnt be here had it not been for their hard work.

Everyone in creative machines lab has immensely contributed to my professional and personal life. I appreciate all the friendship, encouragement, help, collaboration, good ideas, advice, and most importantly, all the laughter from all of them. I will never forget all the happy moments we shared.

Lastly, I want to express my gratitude to my parents, who are an ocean away in China, but who are always just a phone call away. Even though it is difficult for them to understand my line of work since they have never received higher education, their love and support is overwhelming. I

am grateful to them for the freedom that they have given me.

To my mentor, lab mates, and parents

Sincerely, Siyuan Chen

Introduction

The work for this dissertation consists of five parts, with a focus on inferring the causal relationship between time series data. These five parts include Chapter 1, which consists of the basic introduction of concepts, and the other four chapters of research that uncover the causality in various data streams. From earthquakes to financial time series, from mice brain signals to human brain signals. The research of the present study includes two important elements in machine learning: time series data and causality.

As Heraclitus once said, No man ever steps in the same river twice, for it is not the same river and he is not the same man.[121] Nothing is always static. In fact, every stream of data is dynamic. To capture change and temporal memory of time series data is important in understanding the dynamic of the system that produces the data stream. Today, within machine learning, which is the backbone of artificial intelligence, correlation measures are used to discern statistical relationships between observed variables in almost all branches of data-driven scientific inquiry. However, what we are interested in is the existence of causal dependence as it grants us, at least partially, the ability to reason. Statistical tests for causality are significantly harder to construct; the difficulty stemming from both philosophical hurdles in making precise the notion of causality and the practical issues of obtaining an operational procedure from a philosophically sound definition. Specifically, designing an efficient causality test that may be carried out in the absence of restrictive presuppositions on the underlying dynamical structure of the data at hand, is non-trivial. Nevertheless, the ability to computationally infer statistical *prima facie* evidence of causal dependence may yield a far more discriminative tool for data analysis compared to the calculation of simple correlations.

Chattopadhyay and Lipson [11] established a method to calculate the first order causality be-

tween two time series ,and further, if one applies this method to pairs of time series, a first order causality network will be formed.

In Chapter 2, we propose a new non-parametric unsupervised framework to infer the **higher** order causality between quantized time series without making any restrictive assumptions and requirements of prior knowledge of the data. Our model employs the notion of Granger causality but in a quantitative way, which allows us to compare different levels of casualties instead of simple binary statistical tests. Furthermore, our approach demonstrates the dynamic structures between inferred models, which help to construct a deeper understanding of the causal relationship. By implementing our model in Global Industry Classification Standard financial market indexes and cryptocurrency trading data, we show that our method captures well the higher order causal relations between financial time series and also produces deeper interpretations of the results.

After we developed a higher order causality network, other paramount time series problems caught our attention- earthquakes. To be able to reliably predict earthquakes is the holy grail of geology, as large magnitude earthquakes are devastating and the importance of understanding the dynamic and causal relationships of earthquakes as a global activity is self-evident. Earthquake data, as very sparse and highly biased time series, is not easy to process. Essentially, it requires identifying extreme events(large magnitude) out of biased data sets which is dominated by ordinary events(no activities). We examined how large magnitude earthquakes would possibly interact with each others from a data-driven perspective. The present research provides this perspective in the hopes that further research can better predict future earthquakes.

We confirm long-range and long-delay causality effect between seismic events in Middle America Trench and California regions from a data-driven perspective in Chapter 3. Statistical causality was determined using probabilistic finite state automata on historical catalogs of those regions, without recourse to any prior assumption or geophysical knowledge. Pairs of these probabilistic models, with different time delays, were used to assess Granger causality between time histories recorded in these two areas. The analysis revealed that seismic activities of these two regions are linked by around 6 year delay. This finding has an ROC area of 0.62, a value unlikely to happen by

chance, and a confidence level of 99% calculated when compared to coarsely shuffled data streams. These findings corroborate the early conjectures [36] regarding a long-delay relationships between these two particular areas. We suggest that such an analysis can complement geophysical models, and by repeating this analysis exhaustively over the entire planet, we can find a causal network of additional hidden delayed relationships.

Aside from inferring the causal relationship between earthquakes, the same modeling method can be further developed and applied to classify chaotic time series **without** requiring prior knowledge. Chattopadhyay and Lipson [13] used this classification method to differentiate the epileptic seizures signals from normal ones, detecting heart murmurs etc. There is no reason to stop us from applying this method to other signals, such as braincomputer interface(BCI) and all related work is compiled in Chapter3.

BCI is not democratic. Years of training and field research are needed to achieve the a high level of expertise using even the most popular algorithms, which also often requires multiple streams of neural data. We demonstrate how both expert and novice users can use the data-smashing algorithm to classify neural signals of different kinds. To demonstrate, we apply this unsupervised method to distinguish index finger versus thumb movements from the data from a single ECoG electrode, and cluster mice brain activity levels from the Allen Brain Observatory. This is proof-of-principle of a new era in BCI wherein users with minimal field-knowledge and single/modest data streams can obtain classification results that meet and exceed the accuracy of expert users using algorithms that require predefined features, large datasets for training, or multiple data streams.

After processing numerous time series data, figuring out their causal relationship, and dealing with classification problems, we discovered that the temporal pattern captured and used in the previous problems can also be used to condense the information in time series; thus, further helping to compress the data. Subsequently, an index is created using the causal pattern to indicate the predictability of one time series.

In Chapter 5, we propose a new non-parametric distribution-free framework to estimate the time series dataset predictivity without making any restrictive assumptions or needing prior knowl-

edge of the data. Our model is based on previous work called data smashing to calculate the similarities between two time series. By implementing our model in synthetic and Global Industry Classification Standard financial market indexes data,we show that our method performs better than other predictivity estimators because of its ability to capture the temporal memory pattern in the time series.

Chapter 1: Thesis related Basic concepts

1.1 Time Series

A time series is a series of data points recorded in a timely manner. Most commonly, a time series is a sequence of discrete-time data taken at successive fixed length time steps. Examples of time series are open-close-high-low value of equity prices, earthquake actives in the past years, and biology signals.

1.2 Granger Causality

Granger causality is a statistical concept of causality that is based on prediction. According to Granger causality, if a time series X1 "Granger-causes" (or "G-causes") a time series X2, then the past information of X1 should contain information that helps predict X2 without using the past information of X2.

1.3 Data Quantization

In digital signal processing, quantization is the process of mapping input values from a continuous signal to a discrete signal, with a finite number of elements.

1.4 Receiver Operating Characteristics and Area Under Curve

A receiver operating characteristic curve, or ROC curve, is a graphical illustration of the diagnostic ability of a binary classifier system as it changed its discrimination threshold.

The ROC curve is constructed by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall

or probability of detection in machine learning. The false-positive rate is also known as the fall-out or probability of false alarm.

The area under the curve(AUC) is equal to the probability that algorithms correctly classify a randomly chosen positive instance as higher than a randomly chosen negative one.

1.5 Cardinality

In mathematics, the cardinality means a measure of the number of elements of the set. For example, the set Ø={2,-1,70,0.2}. Ø={2,-1,70,0.2} contains 4 elements, and therefore set Ø has a cardinality of 4.

1.6 PFSA

A probabilistic automata, syntactically, is a directed graph whose edges are labeled alphabetically, and the associated transition probabilities. A probabilistic automata with a finite number of nodes (in its minimal description) is a Probabilistic Finite State Automata (PFSA). Formal definition please refer to [21].

1. Identification of -synchronizing string x0: Construct a derivative heap using the observed traces with maximal length set as log(1/) where is the cardinality of the alphabet. For example, for a trinary string (=3), and set as 0.12, then the traces maximal length is calculated as log3(1/0.12)=2.

   The observed traces within this length are null, 0, 1, 00, 01, 10, 11. A derivative heap[21] is the set of probability distributions over cardinality for the subset of strings, which are the observed traces. For example, we calculate the traces tail 0/1 frequency ratio of all the observed traces within the length of 2 in the previous example: null,0,1,00,01,10,11.

   Then we use all the tail frequency ratios to approximate probability distributions by treating the set of frequency ratios as the derivative heap. We then identify a vertex of the convex

hull for the heap, via any standard algorithm for computing the hull. Choose x0 as the -synchronizing string to map to this vertex.

2. Identification of the transition function: We generate transition functions as follows:

   (a) Initialize the set Q as the state q0, set x0 as q0s string identifier and its tail frequency ratio as state probability distribution.

   (b) Then we create a tree structure and set q0 as the root and compute for each trace from that roots tail frequency ratio. If a state q exists, the uniform norm of its and the existing states probability distribution is smaller than , then we merge the state q into the existing state, if not, we define q as a new state.

   (c) The process terminates when a new state is no longer being defined anymore.

   (d) Then, if necessary, we ensure strong connectivity using Tarjans algorithm

1.7 XPFSA

A crossed automata has an input alphabet and an output alphabet, a crossed automata model a finite state probabilistic transducer, that maps strings over the input alphabet to a distribution over set of finite strings over the output alphabet. It is important to remeber that these alphabets need not be identical with respect to their elements or their cardinality. Formal definition please refer to [11].

We perform two steps to train XPFSA which infers the strongly connected minimal realization (a detailed description of the algorithm can be found in [11]):

1. Identification of -synchronizing string x0: Construct a cross-derivative heap using the observed traces with maximal length set as log(1/) where is the cardinality of the alphabet. For example, for a binary string (=2), and set as 0.05, then the traces maximal length is calculated as log21/0.17=4. The observed traces within this length are null, 0, 1, 00, 01, 10, 11, 000, 001, 010, 011, 100, 101, 110, 111, 0000, 0001, 0010, 0011, 0100,

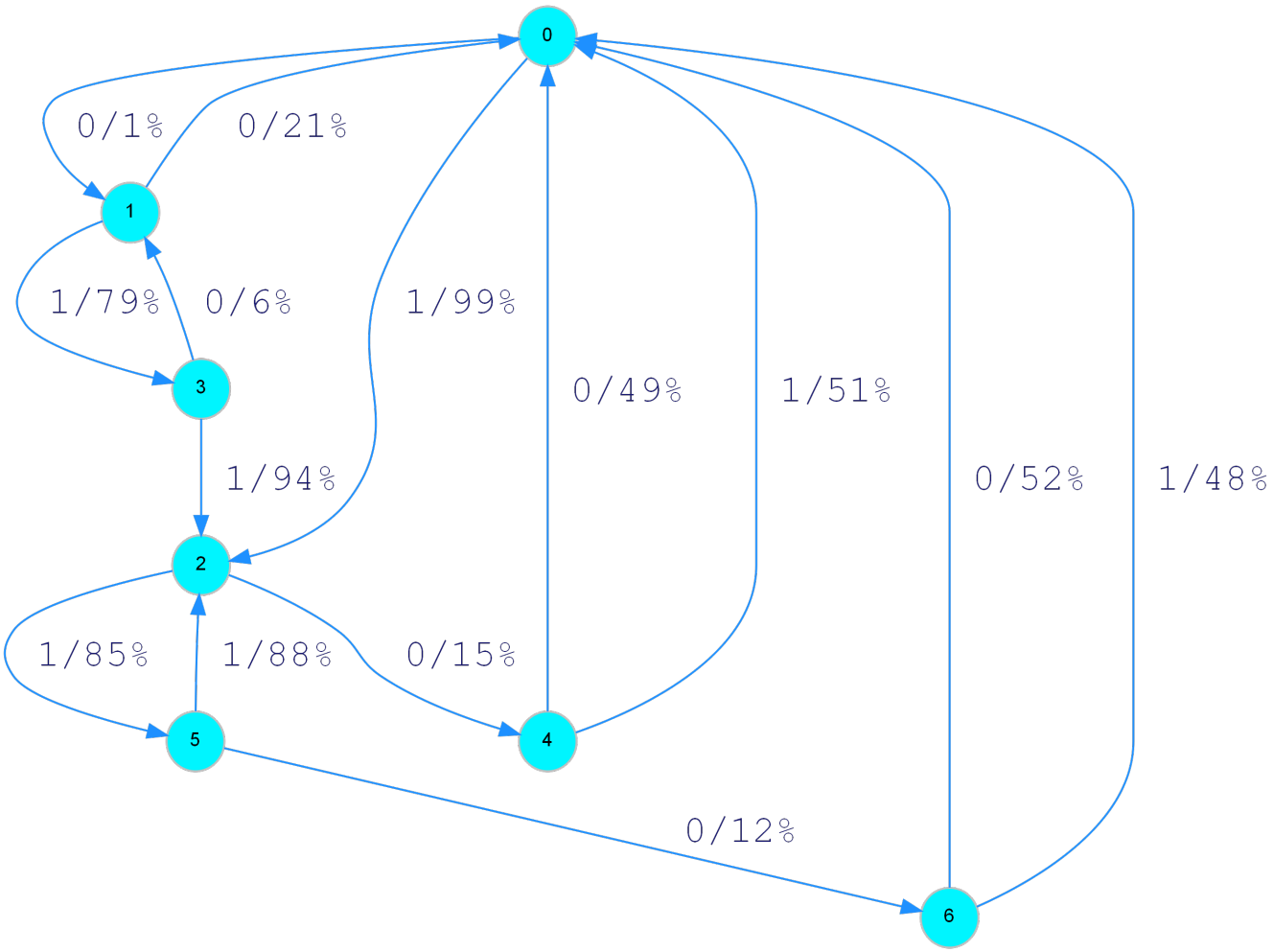


Figure 1.1: An example of PFSA from Chapter 2, Figure. 2.5.

0101, 0110, 0111, 1000, 1001, 1010, 1011, 1100, 1101, 1110, 1111. A cross-derivative heap is the set of probability distributions over cardinality for the subset of strings, which are the observed traces. For example, we calculate the trace Bs tail 0/1 frequency ratio of all the observed traces A within the length of 4 in the previous example, the strings are: null,0,1,00,01,10,11,000,001,010,011,100,101,110,111,0000, 0001, 0010, 0011, 0100, 0101, 0110, 0111, 1000, 1001, 1010, 1011, 1100, 1101, 1110, 1111. Then we use all tail frequency ratios to approximate the probability distributions by treating the set of frequency ratios as a cross-derivative heap. We then identify a vertex of the convex hull for the heap, via any standard algorithm for computing the hull. Choose x0 as the -synchronizing string mapping to this vertex.

2. Identification of the transition function: We generate transition functions as follows:

   (a) Initialize the set Q as state q0, set the x0 as the q0s string identifier and its tail frequency ratio as the state probability distribution.

   (b) Then we create a tree structure and set q0 as the root and compute for each trace from that roots tail frequency ratio. If a state q exists, the uniform norm of its and existing states probability distribution is smaller than , then we merge the state q into an existing state, if not, we define the q as a new state.

   (c) The process terminates when a new state is no longer being defined.

   (d) Then, if necessary, we ensure strong connectivity using Tarjans algorithm

1.8 Data Smashing

Data smashing involves two data streams and includes three steps (see Figure 1): First, raw data streams are quantized by converting the continuous value to a string of characters or symbols. The simplest example of such quantization is where all positive values are mapped to the symbol 1 and all negative values to 0, thus generating a string of bits. Next, we select one of the quantized input streams and generate its anti-stream. Finally, we smash this anti-stream against the remaining quantized input stream and measure what information remains. The remaining information is estimated from the deviation of the resultant stream from flat white noise (FWN). For the formal definition, please refer to [13].

Data-smashing algorithm

The following briefly describes the data smashing procedure to calculate the similarity distance between a pair of signals.

1. *Raw signal quantization*

   Continuous raw signals will be quantized into discrete stream of symbols. In this specific case, we mapped all the values above the 50th percentile of raw signals as ‘1 and the rest as ‘0.

2. *Identify the anti-stream*

   Algorithmically invert one of the quantized streams into its anti-stream which contains the ‘opposite statistical information.

3. *Smash quantized signal with its anti-stream*

   Each stream will be ‘smashed to the anti-streams of all other streams so that their similar statistical structure will be cancelled, leaving the statistical difference only, which will be calculated as a number bounded between 0 and 1 (this represents the probabilistic distance or dissimilarity).

   It is important to note that since data-smashing is an unsupervised method, it does not need the data to be labeled. The pictorial depiction of the data smashing process is shown in Figure. 1.3 and more detailed information on the Data-smashing algorithm is provided in [13].

If the inverted copy of one stream can annihilate the statistical information contained in the other then we can claim that two sets of time series have the same underlying generative process without explicitly knowing or constructing the models themselves. In [13], this property is called information annihilation.

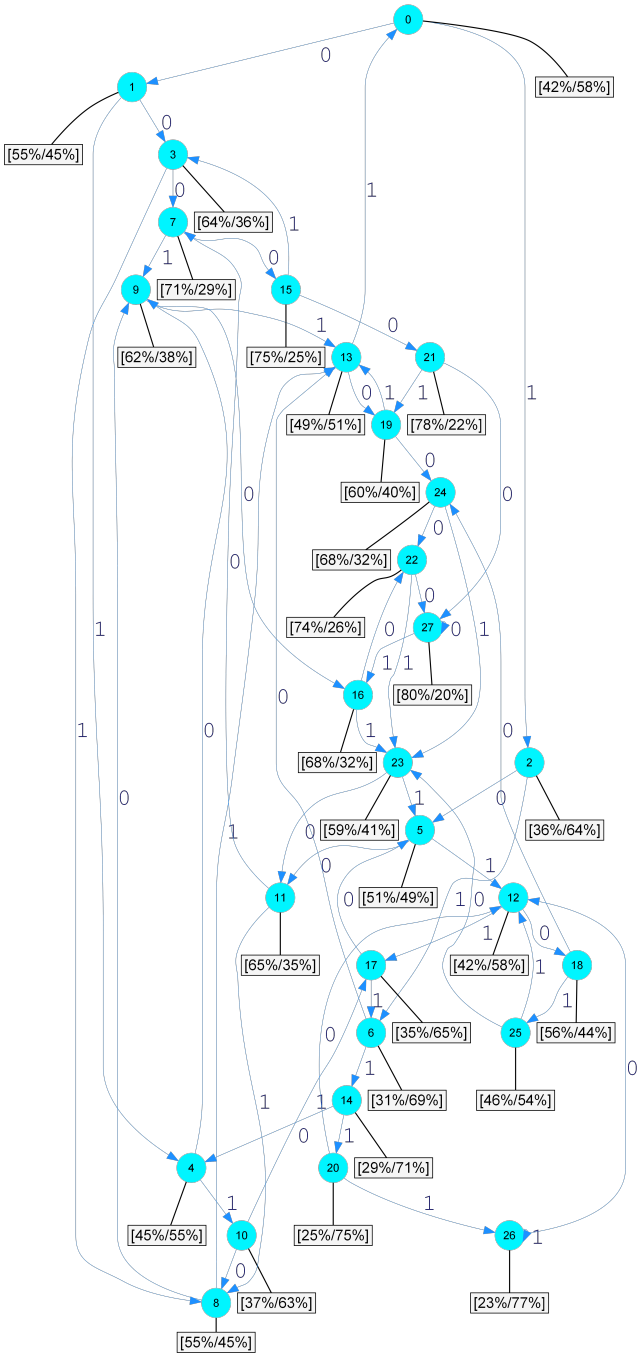


Figure 1.2: An example from Chapter 1, Figure. 2.9

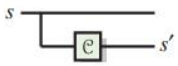
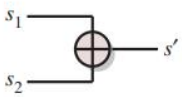
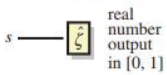
stream operation	algorithmic procedure (pseudocode)
independent stream copy[a]  generate an independent sample path from the same hidden stochastic source	(1) generate stream ω_0 from FWN (2) read current symbol σ_1 from s_1, and σ_2 from ω_0 (3) if $\sigma_1 = \sigma_2$, then write σ_1 to output s' (4) read next symbol and go to step 1 *this operation is required internally in stream inversion*
stream inversion[a]  generate sample path from inverse model of hidden source	(1) generate $\lvert\Sigma\rvert - 1$ independent copies of s_1: $s_1, \cdots, s_{\lvert\Sigma\rvert-1}$ (2) read current symbols σ_i from s_i $(i = 1, \cdots, \lvert\Sigma\rvert - 1)$ (3) if $\sigma_i \neq \sigma_j$ for all distinct i, j, then write $\Sigma \setminus \bigcup_{i=1}^{\lvert\Sigma\rvert-1} \sigma_i$ to output s' (4) read next symbol and go to step 1
stream summation[a]  generating sample path from sum of hidden sources	(1) read current symbols σ_i from s_i $(i = 1, 2)$ (2) if $\sigma_1 = \sigma_2$, then write to output s' (3) read next symbol and go to step 1
deviation from FWN[b]  estimating the deviation of a symbolic stream from FWN (symbolic derivatives (electronic supplementary material, Definition S-9) in the electronic supplementary material, Section S-B, formalize $\phi^s(\cdot)$. If s is generated by a FWN process, then $\phi^s(x) \to \mathcal{U}_\Sigma$ for any $x \in P\Sigma^\star$, and hence $\hat{\zeta}(s, \ell) \to 0$)	$\hat{\zeta}(s, \ell) = \frac{\lvert\Sigma\rvert - 1}{\lvert\Sigma\rvert} \sum_{x:\lvert x\rvert \leqq \ell} \frac{\lVert\phi^s(x) - \mathcal{U}_\Sigma\rVert_\infty}{\lvert\Sigma\rvert^{2\lvert x\rvert}}$, where — $\lvert\Sigma\rvert$ is alphabet size, $\lvert x\rvert$ is the length of string x — ℓ is the maximum length of strings up to which the sum is evaluated. For a given $\epsilon^\star$, we choose $\ell = \ln(1/\epsilon^\star)/\ln(\lvert\Sigma\rvert)$ (see the electronic supplementary material, Proposition SI-15) — $\mathcal{U}_\Sigma$: uniform probability vector of length $\lvert\Sigma\rvert$ — for $\sigma_i \in \Sigma$, $\phi^s(x)\rvert_i = \frac{\text{number of occurrences of } x\sigma_i \text{ in string } s}{\text{number of occurrences of } x \text{ in string } s}$

Figure 1.3: A detailed Algorithms for Data Smashing Stream operations, Chattopadhyay, Ishanu, and Hod Lipson. Reproduced from [22]

Chapter 2: Higher Order Causality and Causality Network

Numerous statistical tests are derived to examine whether correlation exists between variables, as well as to what degree they are correlated. Naturally, the next question we need to ask is whether or not this correlation implies causation' [1]. How do we test for causality?

Granger causality is a well-established statistical concept for describing causality based on prediction. One variable X1 is said to Granger-cause or G-cause another variable X2 if and only if past values of X1 contain information that helps to predict X2 above and beyond the information contained in past values of X2 alone [2]. Many past studies have explored first-order Granger causality and the Granger causality between multiple time series [3,4,5,6,7]. However, only a few have examined higher order causality relationships [8,9].

There are two ways to define higher order causality. One is by looking at the higher order statistical moments, for example, whether one signal can predict the variance or volatility of another signal, rather than predicting the signal itself. This definition of higher order causation is still only applicable only to pairs of data streams (or between a data stream and its own past).

A second definition of higher causation order is to look at causation between three or more streams. Here, we define second order causation as the degree to which one data stream affects the causation between two other data streams.

Prior work mostly focused on variance prediction based on variance of the original signal. By utilizing the Granger causality definition, one variable X1 second-order Granger-causes the other, if X1 conditional variances past information can facilitate predicting X2s conditional variance in the future [10]. Many previous studies employ autoregressive models that require an assumption of the order of the model, parameter estimation etc. Moreover, no higher order moments beyond second order have been explored in these frameworks. Skewness (third), kurtosis (fourth), as well as higher degrees all possess clear statistical meaning and interpretations.

Here, we focus on the latter definition of higher order causation which is essentially causation of causation. We present a novel non-parametric approach to construct generalized probabilistic automata without a *priori* knowledge of the structure and parameters [11]. In addition, we show that the causality network with a higher order can be inferred.

This chapter is organized as follows. Sections 2 and 3 reviewed the past literature, and provide motivation. An overview and discussion of our approach is explained in Sections 4 and 5. Section 6 summarizes the results. Final conclusions are provided in Section 7.

2.1 Related Work

In this section, previous works on higher order causality are reviewed. Almost 20 years ago, Cheung et al [10] developed a two-stage procedure to test causality in variance. The paper estimates the parameters of univariate time-series models and then uses cross-correlation function to test noncausality in variance. Woniak et al [8] proposed a condition for second-order Granger non-causality of a family of GARCH models, and a Bayesian testing procedure of the conditions for Granger non-causality in the conditional mean and non-causality in the conditional variance processes. The generalized autoregressive conditional heteroskedasticity (GARCH) process, which was developed in 1982 by Robert F. Engle, an economist and 2003 winner of the Nobel Memorial Prize for Economics, is used to estimate the volatility in financial markets[8].

In [8], fourth-order moments of time series are assumed. In [9], the author focused on the relationship between causality in mean and causality in variance by using Monte Carlo simulations in the context of Vector autoregressive models (VAR). Furthermore, in [12] [18] [19], the authors extend the notion to a multivariate GARCH model and link it to strong, semi-strong and weak GARCH processes.

Parametric restrictions are derived as conditions to analyze system variables. [18] also proposed a Bayesian testing procedure to evaluate the non-causality hypothesis. However, the assumption that higher-order moments of processes is present is still a requirement. By using one type of multivariate GARCH model and exploiting causality in variance of the source, [20] tried to

explain the generating process of the EEG/MEG signals. Most of the paper focused only on studying the variance the second order causality between time series, and they are almost all parametric and model-based.

2.2 Motivation

Chattopadhyay and Lipson[12] proposed a nonlinear (non-parametric) modeling approach using PFSAs to calculate self-causal relationships for ergodic and stationary quantized stochastic processes. At least one contribution of their work [11,12] explore what insight a generative nonlinear model of self and cross dependence that was distilled from the data at hand can offer.

Specifically, Chattopadhyay and Lipson[12] offered a quantitative way of measuring the degree of causal influence that exceeds simple binary hypothesis testing, without any prior knowledge and presumption of model structure. In the paper, the degree of causation is referred to as a coefficient for causal dependence in the paper. Before that work [11,12], most of the state of the art techniques, such as the Hiemstra-Jones (HJ) test [13] which is also a non-parametric approach, have been set up in a way to detect the existence or non-existance of a causal relationship, i.e., the framework of a classical binary hypothesis testing.

The quantitative way provided by Chattopadhyay and Lipson[12] shed light on a way to infer the different degrees of a causal relationship between a pair of data streams using stochastic modeling.

The comparison between the different degrees of causal dependence generated by the different dynamic causal structures inspired a new signal classification algorithm called data smashing [13], which allows us do classification in a plug and play style without any domain knowledge and data preprocessing.

It is natural to attempt to extend this idea to second and higher orders. The aim of this paper is to set up the framework for second order causality inference and test our approach by using a toy problem and two real world datasets. We focus on econometrics as our primary application.

Various kinds of Granger causality tests have been proven successful in econometrics, includ-

ing the discovery of nonlinear causality between stock returns and macroeconomic factors [15], income and money [1414], currency future returns [16] and stock price and trading volume [17]. It would also be also very interesting to explore how second order causality networks can be applied to financial domains and what deeper intuition arises through inferred directional second order.

2.3 Higher Order Causality Networks

Here, we present two different but similar process to infer higher order causal relationships for self and cross-quantized stochastic processes.

1. For cross-causality inference:

   (a) Given two quantized time series: Driving string A, driven string B. Implement GenESesSS[11] to infer strongly connected XPFSA. There are several machine learning models that can model the stochastic process. We chose to use a XPFSA because it is proven to be able to capture the long-term causal relationship from a driving string A to driven string B [11,21]. For details, please refer to Chapter 1, Section 1.7.

   (b) Rerun the XPFSA graph model by reading the driving string A and record the states being visited during the process.

   (c) Given the trace of states, read off each states corresponding probability distribution. For each state, assign it with a symbol with the highest probability. For example, for a binary string, the uniform distribution of the probability of two symbols appear in the system is 50% in the system. Assign 0 to the state whose 0/1 probability ratio is larger or equal than 1, and assign 1 if the 0/1 ratio is smaller than 1. Here, we call it the ***Hidden Causal State trace string***is written as ***(A->B)***.

   (d) Given another string C and D, implement xGenESesSS again to infer strongly connected XPFSA, then we get Hidden Causal State trace string (C->D).

   (e) Implement xGenESesSS again to infer strongly connected XPFSA of (A->B) and (C->D), which leads us to the second order causality relationship.

(f) Calculate coefficient of causality (a detailed description of the algorithm can be found in [11])

(g) Repeat the process as shown in Figure. 2.1 to infer higher cross order causality.

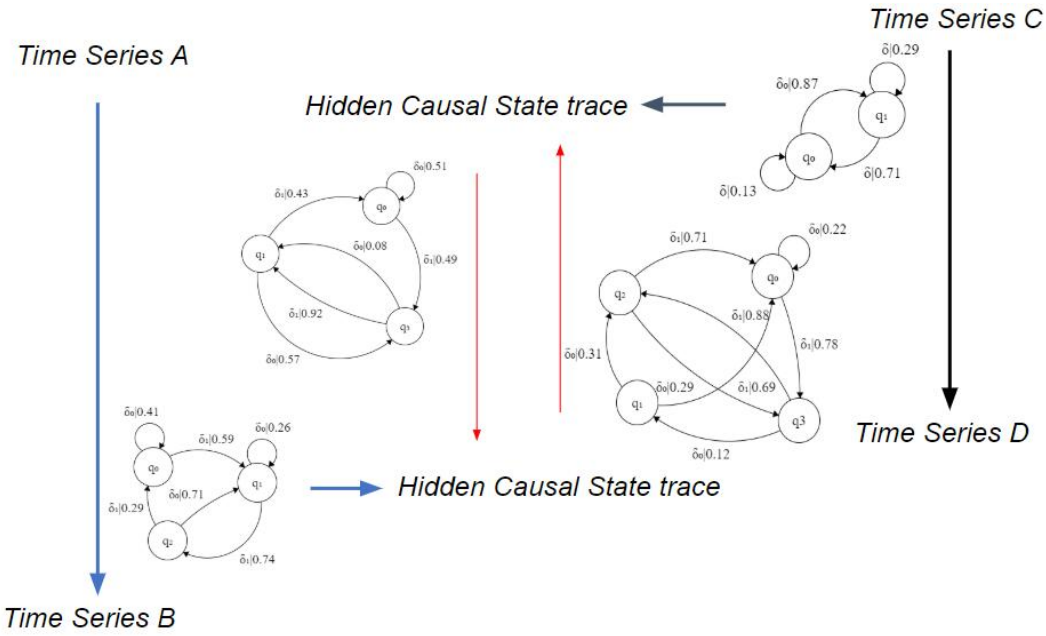


Figure 2.1: Paradigm of how one layer of higher order causality is calculated, time series A,B,C,D are quantized streams. This example tries infers the causality between the causality of AB and CD

2. For self-causality inference:

(a) Given one quantized time series: Implement GenESeSS [21] to infer strongly connected PFSA. PFSA, a slight different version of XPFSA, is also proven to capture the long-term causal relationship from one time series past information to its future information. We perform two steps to train PFSA which infers the strongly connected minimal realization (a detailed description of the algorithm can be found in [21]):

(b) Rerun the PFSA graph model by reading string A and record the states being visited during the process.

(c) Given the trace of states, read off each states corresponding probability distribution. For each state, assign it with the symbol with the highest probability. Here, the Hidden Causal State trace string would be(A->A).

(d) Implement GenESesSS again to infer strongly connected PFSA of (A->A). This leads us to second order causality relationship.

(e) Calculate coefficient of causality [11]. Repeating the process will lead to higher self-order causality.

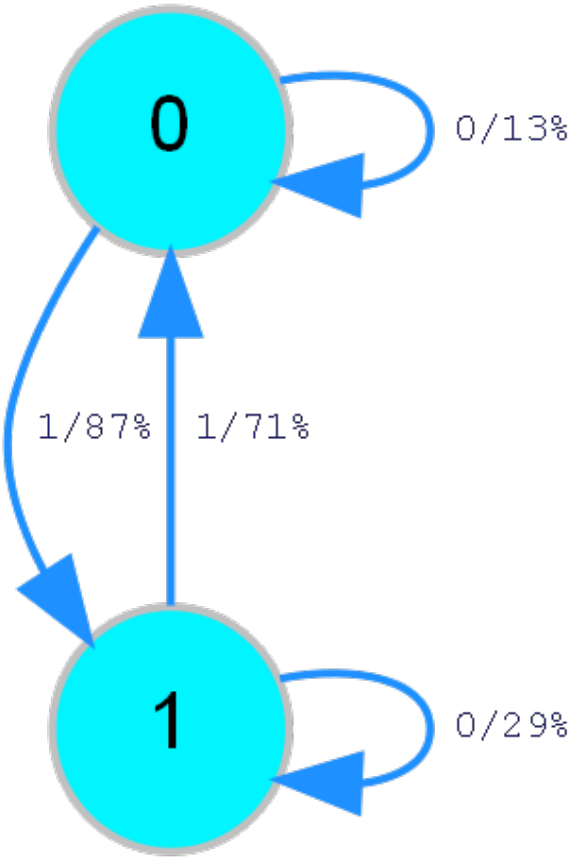


Figure 2.2: An simple example of an inferred diagram. The state machine has two states and directional connections and the associated weights are shown in the figure.

2.4 Discussion

The intuition behind the Fig. 2.2 is as follows: If one given two binary time series A and B, and As past information has a direct causal influence on future B and is governed by dynamical structure shown in figure 2.2. The XPFSA inferred has two states, if the state triggered by time series A hits Q0, this means B has probability ratio of 87/13 to generate 0/1, or if Q1, B has

probability ratio of 29/71 to generate 0/1. Baseline would be 50/50 if no extra external information is provided.

Given the inferred generative XPFSA model, the systems causal dynamical structure is revealed, entropy rate for stochastic process gets lower.

The idea behind higher order causal relationship is as follows: Instead of calculating XPFSA, we can calculate the hidden causal state time series (A->B) inferred from raw data A and B, then calculate (C->D). XPFSA between time series (A->B) and (C->D) can then be inferred. We then we consider this process one layer deeper into the causal relationship as shown in Figure. 2.1.

The XPFSA inferred from (A->B) and (C->D) is the causal relationship of the causal relationship between A, B, C, D. For (A->B). This can be considered as a representation of second order moment, an analog of variance between A and B.

Take Figure. 2.2 XPFSA as an example. The ratio of 0/1 at different states (Q0&Q1) are 87/13 and 29/71. The mean of two states distribution is much closer to 50/50 compared to each individual state. In this perspective, the string recording the change of state, to a certain extent, is perhaps related to the variance because it shows how the causal probability fluctuates with respect to time.

This is a unique directional causal variance between A and B. It is analogous but different compare to variance. It involved two time series, and not like co-variance which is non-directional (COV (X, Y) = E[(X-E[X]) (Y-E[Y])]), causal relationship is directional.

As for a time series self-causation problem, string resulting from (A->A) reflects the fluctuation of the state causal probability of time series itself. This encodes information of the deviation from its mean, which is a concept that is conceptually close to variance but in a graphic model approach.

We can further to explore higher order causality by looping the process: Take time series self-causal string as a function of time to get (A->A). The PFSA of (A->A) can be calculated and then its causal state trace string ((A->A)->A)=(A->->A) can also be calculated. (A->->A) can be viewed as the skewness which describes the asymmetry of the causal probability distribution.

And moving further forward to infer (A->->->A), a concept similar to kurtosis, we can achieve by repeating the same procedure after (A->->A).

A very similar but different model would be the hierarchical hidden Markov model(HHMM). In the HHMM, each state is a self-contained probabilistic model. More precisely, each state of the HHMM is, itself, an HHMM [22]. When calculating causal relationships in our approach, the underlying transition states of state machine are treated as time series again, and the PFSA/XPFSA is then calculated. Our inferred automata are fully connected so that the entropy on of change can be measured and the causal relationship can be inferred.

2.5 Experiments and Applications

Our experiments and applications are conducted on three different types of time series datasets: simulated data, daily industrial sector price time series data of the Global Industry Classification Standard (GICS) level I, II sectors data, and intra-day 10-seconds bitcoin spot prices.

For every time series of the industrial sector price, the data begins on January 2, 2002 and ends on July 14, 2017. Each time series has 16,380 observations. GICS is a tiered, hierarchical industry classification system, where companies are classified quantitatively and qualitatively in the system.

Each different level represents one tier of industry level; the higher the tier, the more detailed the data in that sublevel. For example, level one includes Technology, Energy etc. and level three includes more detailed Energy Equipment Services, Oil, Gas Consumable Fuels. For the intra-day Bitcoin spot price, January 2018 at 10-second prices are used and each series has around 374,400 observations.

2.5.1 Toy Problem

Here, we present a toy problem for second order cross causality inference. The purpose of the simulation experiment is to verify whether our method can infer the true causal relationship of the conditional distribution of the given time series.

We use four binary time series labeled A, B, C, D, whose first order XPFSA machines and second order XPFSA machines are a predefined known prior. Their XPFSAs connection, states and arc distribution were pre-designed, and we inferred predefined causality as outputs.

In total, 1000 data points are generated for each data stream. We then compare the results of causality of inferring four complete randomly generated time series.

After repeating the simulation for 1000 times, we plot three histograms which stands for inferred and random second order causality.

As shown in the Figure. 2.3, the red vertical line is the designed second order causality, set at 0.15. The orange histogram stands for the inferred the second order causality, where its mean lies very close to the designed level. At the same time, the random generated time series casualties clustering around zero. This control experiment demonstrates the algorithms ability to infer second order causal relationship between time series.

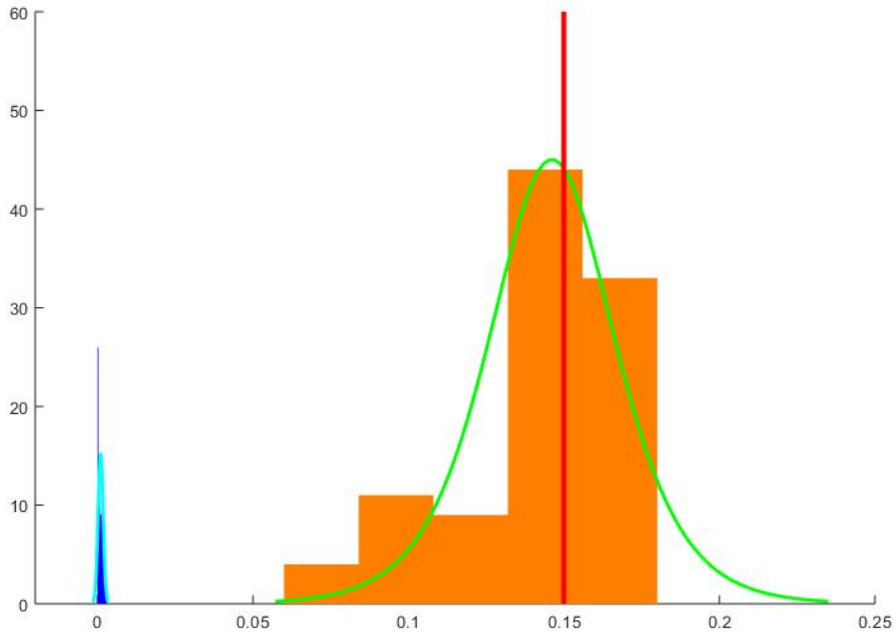


Figure 2.3: The vertical red line is the designed second order causality. The orange histogram is inferred causality for 100 times, and its mean is around designed causality. The green curve is its distribution fit. The left blue peak close to 0 is the distribution of causality inferred from the randomized data stream. The navy blue curve is its distribution fit

2.5.2 Second Order Causality Network of Global Market Data

The Bloomberg terminal [23] provides a convenient API to download the industrial sector price time series data. The time series of prices are calculated and quantized into binary strings, with the symbol 0, indicating a negative movement in return, and a 1, indicating zero or positive movement in return.

It would be interesting to study the sector datas second moments (also known as volatility) causal relationship instead of correlation. We first calculated each time series' self-hidden causal state trace string, then constructed the second order causality using a cross model.

Figure. 2.4 shows the self and cross coefficient of causality between all sectors in level I data. The first column is the driving stream and first the row is the driven stream. For example, the industrial next day second-order causal influence to Technology is 1.667648559*e-3, which is the third element under the Technology column. It can be seen that some of the strong next day second-order causality lies on the diagonal, for example, Technology, Energy, Healthcare. Similar to Technology, its self-coefficient of causality is 121.96*e-3 (highest in Figure. 2.4). Its PFSA, as shown in Figure. 2.5, has 7 states with most of the states exhibiting high causal relationships, meaning that Technology has a stronger self-temporal memory pattern of their volatility compared to any other sectors calculated.

× 0.001	Technology	Consumer Discretionary	Industrials	Basic Materials	Financials	Real Estate	Energy	Healthcare	Consumer Staples	Utilities	Telecom Services
Technology	121.9553972	0	0	0	0	6.69878316	5.497444336	3.203153397	0	7.771400797	1.002785784
Consumer Discretionary	0	0	0	0	0	0.9235435433	0.2828380548	0	3.086435926	3.827297889	3.542952125
Industrials	1.667648559	1.218286112	0	0	2.944958967	1.385076405	1.087845085	5.996325222	18.25899986	0.4768946544	3.127735177
Basic Materials	0	0	0	0	0	0	0	0	0	0	0
Financials	12.66751632	3.009796495	4.41610812	0	31.32612191	0	1.905517823	3.832541663	1.220345291	2.432139487	4.223672823
Real Estate	4.246584997	0.2984191158	7.239906924	0	0	0	3.426394643	4.691746854	0	0.8381456413	2.608453357
Energy	4.525302658	0.643499548	15.05236839	0	2.524063652	0.5467679012	80.64826132	1.98027416	1.157561812	0	4.240163332
Healthcare	7.372457722	5.974013954	7.962313721	0	0.2486933718	1.626222564	5.565876712	37.58140261	0.7358605235	3.959124172	2.810271649
Consumer Staples	1.724450462	0	0	0	0.723173384	0	2.659646578	0.6758343041	0	0	0.7709184937
Utilities	0	5.820357426	5.172025481	0	2.353667995	0.44131097	0.3156485011	6.109837065	5.855020317	0	6.602955586
Telecom Services	4.75103296	4.113201674	6.483998728	0	6.846424284	0.6122605929	3.077270366	6.07587439	13.22930029	3.376479554	19.93848707

Figure 2.4: Level I GICS (11 entries) causality table. The green is highlighted cells represent a causality higher than 10*e-3

Figure. 2.6 shows a second order causality network between level I sectors data. The network shows a strong causal influence from Financials to Technology and Industrials to Consumer Staples. Basic Materials has very little influence on all the other sectors. Its sublevel second order

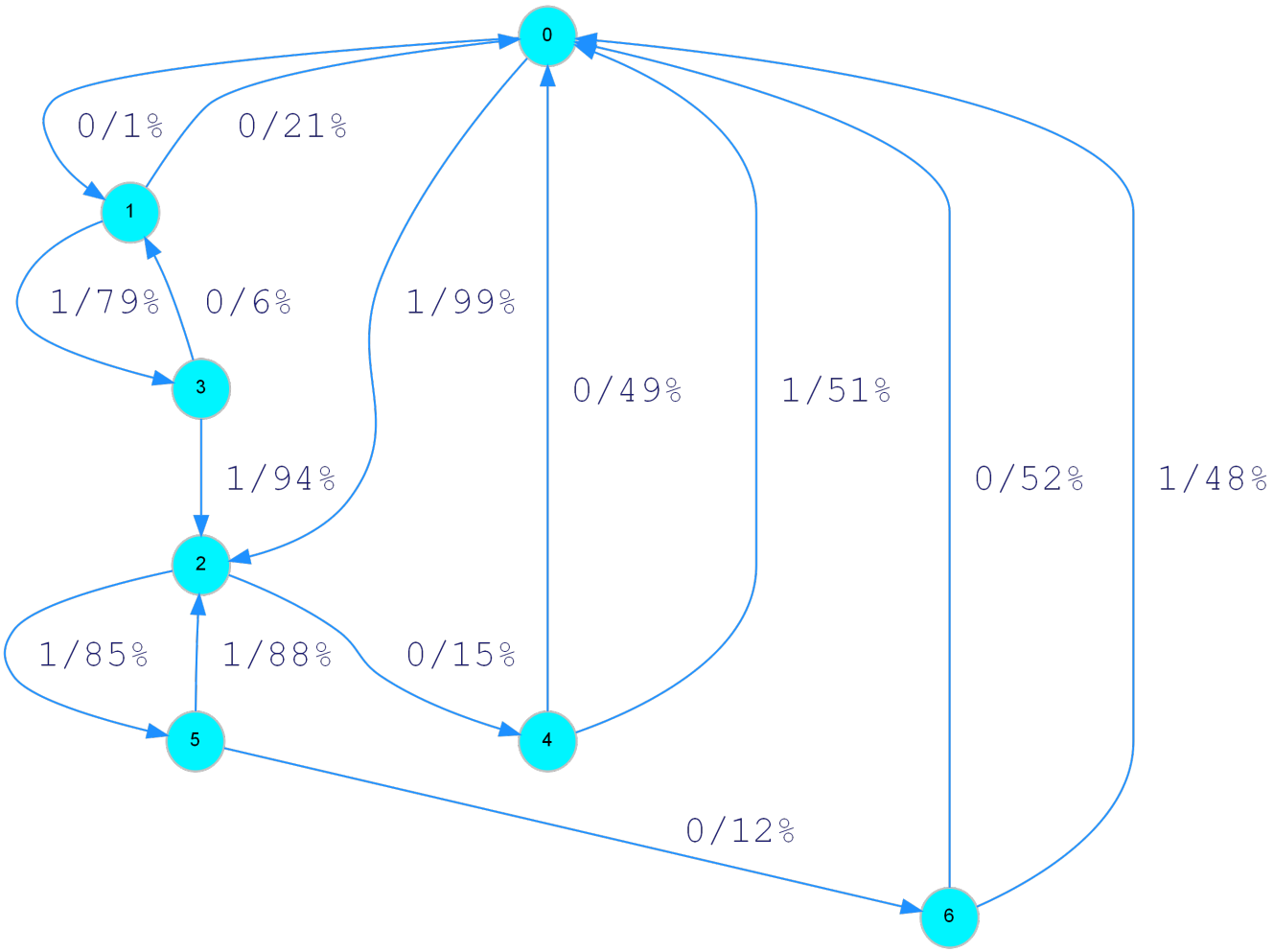


Figure 2.5: Inferred Technology second order causal state machine(self). It has 7 states and most of the states possess low information entropy level [24]. For examples, states with distribution[6%/ 94%],[98%/ 2%]. Their information entropy is 0.2423, 0.1414 respectively, which is much lower than the original Technology information entropy 0.63.

causality network is plotted below.

Figure. 2.7 shows a second order causality network between level II different industrial sectors data. The Materials sector, Automobiles Components, and Software Services sectors show close to zero causal relationship to the others. However, the Household Personal Products sector exhibits a high second order causal relationship to the Transportation sector. This means that when the Household Personal Products sector price is volatile, the Banks sector will experience volatility too. Therefore, traders can lower their bets risk in the Household Personal Product sector by hedging in the Transportation sector.

According to Business Insider, in 2015/08/12, Hedge funds are going long large-cap financial companies and transportation stocks and ***shorting*** energy and household- and personal-product companies. [25]

In addition, in Figure. 2.6, Financials has a strong causal relationship pointing to Technology. According to [26], Diversified Financials and Semiconductors Semiconductor Equipment are their sublevel, respectively, and possess the same direction influence as shown in Figure. 2.7.

Another observation is that in level II, Household Personal Products have a strong influence on Transportation. Meanwhile, their associated level I sectors, Consumer Staple and Industrials, has a strong causal relationship but with a reverse direction.

2.5.3 Bitcoin Trading Arbitrage

Cryptocurrency like bitcoin (BTC), is a form of electronic money operating without a central bank. Money can be sent and received in a peer-to-peer decentralized network without any intermediaries [27]. Among other cryptocurrencies, Bitcoin is largest in terms of its market capitalization among other cryptocurrencies. Cryptocurrency can be traded on crypto exchanges once they are listed [28]. The price of bitcoin hit an all-time-high, close to 20,000 USD per bitcoin, on December,16, 2017. Trading activities in cryptocurrency skyrocketed during that period of time. With the maturity of exchange infrastructure and novel crypto hedge funds, more sophisticated trading strategies made their way to crypto trading. Here, we explore the simplest one which is called arbitrage: a way to trade for profit by simultaneous purchase and sale of an asset from an imbalance in the price. The estimated total amount of arbitrage profits just from December 2017 to February 2018 was above of $1 billion [29]. This was because of lead-lag between different exchanges and substantial variation in the level of liquidity across different exchanges and currency pairs[30]. The leadlag relation between price movements of one exchange and another illustrates how fast one marketplace reflects new information relative to the other, and how well the two exchanges are linked[31]. By looking at the trading data from the exchange, and to calculate the cross causality, arbitrage activity can be detected and explained perfectly by causality inferred.

We capture the price and volume data (in the period 2018.1.1 to 2018.1.26) in two crypto exchanges, Bitmex perpetual swap BTC/USD(A) and Gdax BTC/USD(B), through their API portal. We quantize the volume and price into binary time series, with the symbol 0, indicating a drop in the price or volume, and a 1, indicating zero or an increased in price or volume. We then looked at each exchange price to volume causal relationship, price A to volume A, and price B to volume B.

Once we found each hidden causal state of each trace, we calculated the coefficient of causality from (Price A->Volume A) to (Price B->Volume B) and vice versa. This is the second order causality because the inferred causal relationships are inferred between hidden causal state traces (Fig. 2.11).

All the pairs of the two exchanges price and volume coefficient of casualties are computed and listed in Figure. 2.8. Typically, a change in the movement of the price will attract more trading volume. Hence, why the causal relationship from price to volume is explored first. Arbitrage happens when there is a discrepancy in the price between two exchanges. Often, the price moves first in one exchange and then the other exchanges follow.

As shown in Figure. 2.11, the second order causality, exchange A, is much higher (0.0697) than B (0.0016), which means that Bitmex in the selected period is the price first mover compared to Gdax.

Is this finding supported by first the order causal relationship? The answer is in the affirmative. Figure. 2.8 shows the first order causality table between the Bitmex(A) price and volume and the Gdax(B) price and volume. For the sake of simplicity, in the following paragraph, A is referred to as Bitmex and B is referred to as Gdax.

There are two interesting observations in the first order causal relationships: First, for price only, price(A)s causality (0.0328) on price(B) is higher than price(B) self-causality (0.0316). This shows that As price has a higher influence on Bs price compare to Bs price itself. Second, Volume(A) has a higher causal influence on price(A) (0.0022) than volume(A) to price(B) (0.0016), while volume(B) to Bs price is only 0.0012.

Raising/dropping of the trading volume directly affects the price [32]. However, As volume has

a higher influence on price B than Bs volume to its own price. The interpretation is that changes of volume in A drove the price in A, and this driving force influenced the price in B more than Bs volume itself.

By further examining and calculating the second order causality, our approach captures this phenomenon well and reveals the causal relationship much more clearly than if we had solely focused on the surface of first order causality. The second order causality from A to B is 0.0697 which is much higher than the second order causality from B to A and other first order casualties in magnitude.

Figure. 2.9 shows the complete picture of the second order causality network from Bitmex to Gdax. There there are 28 causal states which represent the causality of causality pointing from the price of Bitmex to volume to price of Gdax to volume. Some of the generated states, for examples, states 15 and 21 (among other low entropy states) are the distinctive states indicate that Bitmex is the major driving force that lead the trading activity in Gdax.

Given this inferred relationship, the temporal memory that drives to states such as 15 can further be processed into the trading signal. however, the network could be changing all the time, and this specific plot is just the historical data from the beginning of 2018. Relationships, such as Figure. 1.10 can be inferred very quickly because of their proven PAC learnability [11,21], which can be very helpful and important in real world applications[33,34,35].

For the sake of completeness, we mapped out a few of the other large volume exchanges second order causality networks as shown in Figure. 2.11. We calculated the second order causality network between Bitmex and Gdax in the same way but with the following exchange pairs: Bitfinex BTC/USD, Bitflyer BTC/JPY, Bitflyer BTC/USDT, Bitmex quartly BTC/USD, Bitstamp BTC/USD, Bittrex BTC/USDT, Coincheck BTC/JPY, Kraken BTC/JPY, Kraken BTC/USD, Okex weekly/bi weekly/quarterly BTC/USD, Poloniex BTC/USDT.

From Figure. 2.10, it is evident that the USD and Bitcoin perpetual swap pair in the Bitmex, USD and Bitcoin quarterly future contract in Okex and JPY and Bitcoin trading pair in Coincheck are the leading trading and price driving force in the market, given that the exchange provideds the

real trading data.

Other exchanges and pairs play a much weaker role in the market, such as Bitflyer, Bitstamp or weekly future contracts in Okex.

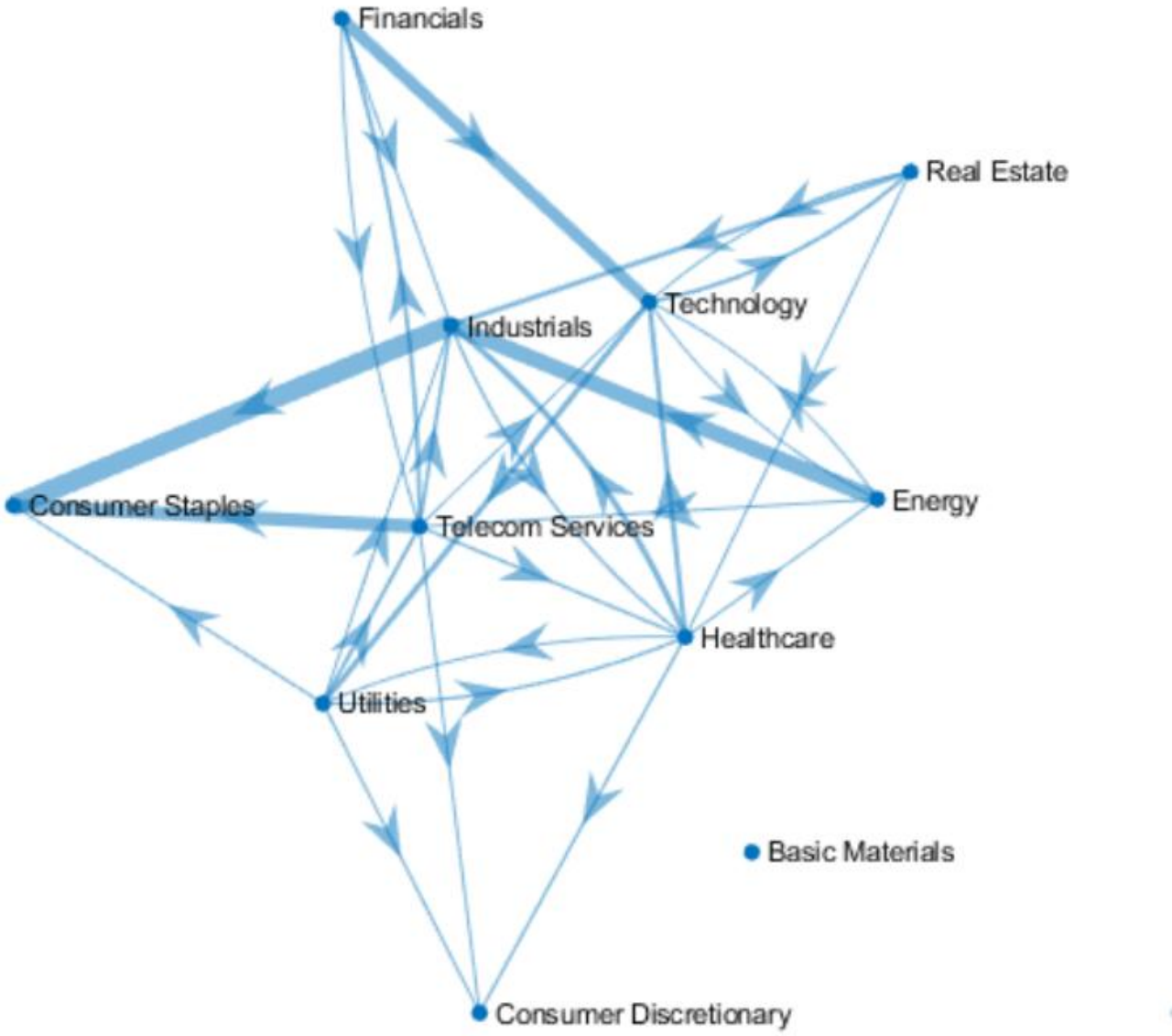


Figure 2.6: Second order causality network of Level I data. Strong linkage can be found: Financial point to Technology, Energy to Industrials, Industrials to Consumer Staples

Figure 2.7: Second order causality network of Level II data. Strong linkage can be found: Diversified Financials point to Semiconductors Semiconductor Equipment, Energy to Semiconductors Semiconductor Equipment, Household Personal Products to Transportation

Causal Coeff	Price(A)	Price(B)	Volume(A)	Volume(B)
Price(A)	0.0064	0.0328	0.0024	0.0003
Price(B)	0.0015	0.0316	0.0008	0.0006
Volume(A)	0.0022	0.0016	0.0019	0.0003
Volume(B)	0.0006	0.0012	0.0002	0.0009

Figure 2.8: Bitmex perpetual swap BTC/USD(A) and Gdax BTC/USD(B)s price and trading volume first order coefficient of causality table

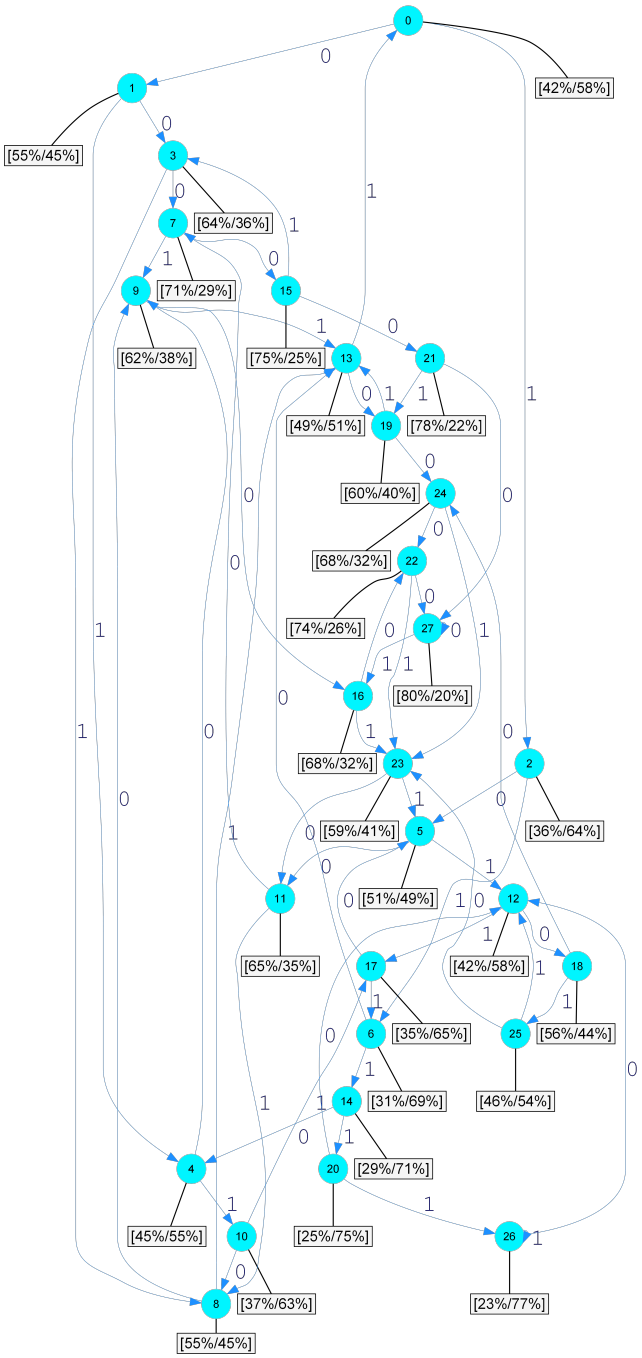


Figure 2.9: Bitmex to Gdax cross second order causality network, with 28 causal states and the states distribution

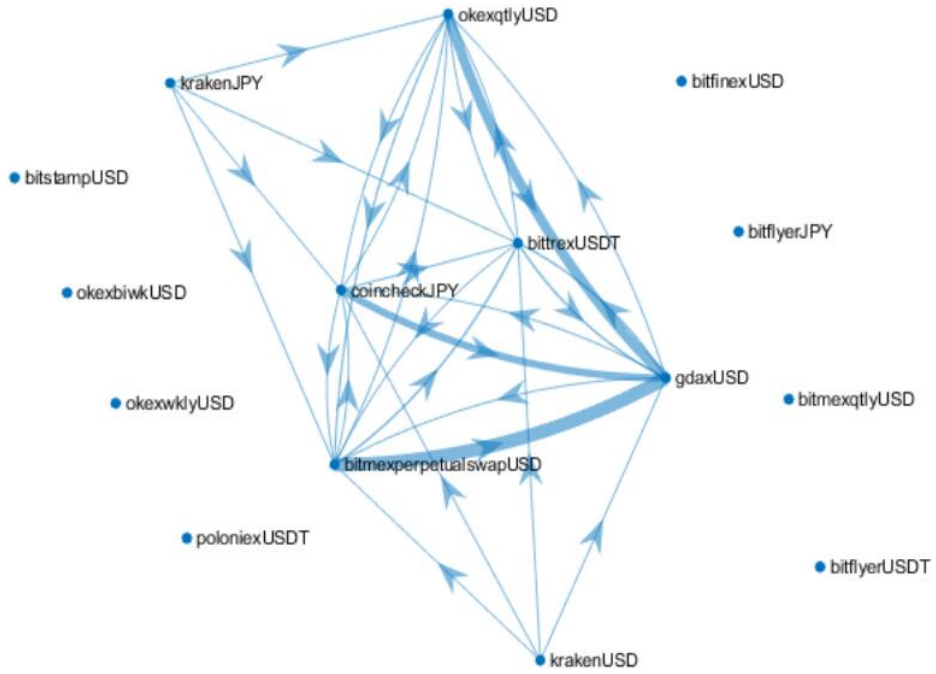


Figure 2.10: More comprehensive second order causality network between different pairs at different exchanges. Bitmex perpetual swap and Okex quarterly future contract manifest themselves as the driving force of the BTC secondary market.

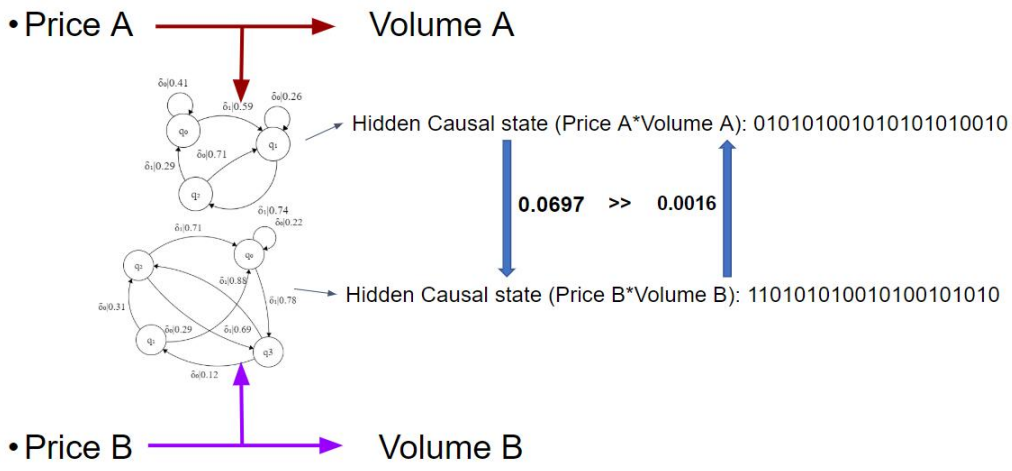


Figure 2.11: Pipeline for calculation of second order causality between Bitmex(A) and Gdax(B), 1). Calculate two first order causality networks from the Bitmex price to the Bitmex Volume and the Gdax price to the Gdax Volume. 2). Calculate the quantized hidden causal traces 3). Calculate the causality network between two quantized hidden causal traces in both directions.

Chapter 3: Causality Pattern between Seismic Activity in Middle America Trench and California

3.1 Introduction

Computational forecasting models for major earthquakes are frequently based on the analysis of the quantitative relationship between the occurrences of significant seismic events at the local spatial scale and the short-term temporal scale. Such approaches date back to the 1970s, before any substantial global data catalogs were available. For example, in 1970, Sadovsky and Nersesov [37] suggested that geophysical parameters, such as the ratio of velocities of longitudinal and shear waves, vary before a strong ground motions across the focal region. In this paper, we use statistical methods to investigate the possibility of a long-distance ($\geq$ 3000 km) and long-time delay ($\geq$ 3 years) causality between seismic events. Here we use the strict Granger definition of causality [38], whereby one time series has a causal relationship to a second time series, only if the first series contains information that reduced the entropy (uncertainty) of the second time series at a future time.

To test such a hypothesis, a new stochastic modeling algorithm was developed and applied to seismic activities associated with Middle America Trench and California, separated by approximately 3038 kilometers. The seismic activities of these two earthquake-prone locations are well recorded [39,40], and found to be evolving in a somewhat synchronized manner, as indicated by Raleigh et al. [36].

Our results indicate that many subtle seismic Spatio-temporal interaction relationships between California and the Middle America Trench indeed exist, and when combined point to an intriguing causal influence among earthquake activities. Specifically, events in the Middle America Trench may somehow influence the activities in California, 3038 km away and 6 years in the future.

While the underlying physics of these long range delayed effects is not yet clearly defined, the combined significance of these newfound relationships is statistically non-trivial and robust, and has significant scientific and practical implications.

Section 2 surveyed the current states of earthquake forecasting. Section 3 presented the data source and modeling approach. The definition of causality was then elucidated in Section 4. Sections 5 and 6 presented the data preprocessing, modeling methods, details and evaluating metric. Control experiments and conclusions are finally expatiated in Sections 7 and 8.

3.2 State of Earthquake Forecasting Research

Seismic events with extended time delays on the order of years are not explored extensively in the literature. Table 3.1 surveys 17 papers on quantitative earthquake predictions and shows that the majority of long-delay studies focus on distances less than 1000 km.

Table 3.1: Review on the past literatures

Approach	Distance (km)	Delay	Causality	Year
Microseismicity and Strain buildup [36]	30-200 [1]	Years-decades	Multiple/Self	1982
Statistical model [42]	0-1200 (km)	0-30 years	Network/Self	1990
Observation on the real-world seismic events [44]	10s-400[2]	14-35 years	Multiple/Self	1990
Statistical model [45]	50	30 years	Single/Self	1993
Statistical model ([46]	300-500	5-25 years	Multiple/Self	1996
Phase Transition theory [47]	325[3]	1.5-80 years	Multiple/Self	1998
Observation on the real-world seismic events [48]	3100-3660[4]	13-15 minutes[5]	Single/Self	2004
Statistical model [49]	0-300	0-3 years	Network/Self	2004
Statistical model [50]	Within California	1 day	Single/Self	2005
Observation on the real-world seismic events [41]	0-60	0-23 years	Multiple/Self	2005
Stress analysis [51]	0-240[6]	0-10 years	Single/Self	2008
Statistical model [52]	30	1 day	Single/Self	2009
Stress analysis [53]	50	20 years	Single/Self	2012
Statistical model [54]	250	5 years	Single/Self	2013
Deep learning [55]	100-300[7]	30 days	Multiple/Self	2018
Stress analysis [56]	116-25[8]	10-22 years	Network/Self	2018
Deep learning [57]	2000-3000	0- 1 year	Multiple/Cross	2018
This paper: Statistical model	2000-3000	0-11.5 years	Single/Cross	2019

Raleigh et al. [36] examined the long-term variation of the occurrence rate of earthquake events, in order to predict upcoming strong shocks in Southern California. The underlying physical assumption associated with their model lies in the hypothesis that earthquakes result from the accumulation of elastic strain in the brittle lithosphere. In addition, they believed strong earthquakes are typically found to be ensued by the slow buildup of strain, as well as a set of shocks with a lower magnitude across that specific area. Considering the increased frequency of moderate to large earthquakes in Southern California over the preceding few years, they surmised that the seismic events with magnitude higher than seven would be likely to happen during the upcoming decade. In retrospect, that forecast, though quite vague, came to fruition, insomuch as the occurrence of the Landers Earthquake with a magnitude of 7.3, in 1992.

Schorlemmer and Wiemer [41] also studied the size distribution of micro-earthquakes recorded during the decades before the main shock with a magnitude of 6.0 at Parkfield, California, 2004. They regarded the unusually low b values (associated with the Gutenberg-Richter law) around that region as an indicator of highly stressed patches in the fault, and therefore argued that such values could be used for predicting ruptured areas.

Nishenko and Bollinger [42] also tried to forecast the next strong shock throughout the Central and Eastern United States, by considering the recurrence pattern of earthquake events, yet from a statistical perspective. As the return interval between strong earthquakes is usually remarkably longer than the time elapsed since the last strong event, the Poisson model (that is time-dependent) offered an approximation of the seismic hazard level. By taking such an approach, they argued that the Poisson probability for an earthquake event with a magnitude higher than 6 over the next 30 years is at a moderate to high (0.4 to 0.6) level for such regions, based on the frequency-magnitude analysis of both the seismograph network and the historic earthquake catalogs. Moreover, accord-

[8](1) The vicinity of the coming rupture of strong earthquakes, which is usually 30-200 km long.
[8](2) Measured based on the location of specific seismic events with intermediate magnitudes.
[8](3) Claimed to the radius of the "*optimal*" critical region.
[8](4) Claimed to be the most distant case of remotely triggered seismicity yet observed.
[8](5) Demonstrated in the manuscript.
[8](6) The dimension of the rupture.
[8](7) The dimension of the area under consideration.
[8](8) Demonstrated in the manuscript.

ing to their model, the corresponding probability would be even higher (0.7 to 0.8), if the entire Eastern North America was considered entirely.

In light of the fast-growing Machine Learning (ML) methods throughout the past decades, some researchers have conducted a set of pilot studies as an effort to apply ML methods to earthquake predictions [66,67,68,69]. Rouet-Leduc et al. [43] applied ML techniques to datasets from shear laboratory experiments, in order to identify the potential hidden signals foreshadowing earthquakes. It was shown that by tracking the acoustic signal emitted by a laboratory fault, the machine learning systems can forecast the time remaining before the eventual failure with statistically significant accuracy. They further argued that such an approach would be capable of spotting unknown signals and placing bounds on fault failure times. Importantly, ML techniques could identify the signal emitted from the fault zone previously thought to be low-amplitude noise that enables failure forecasting throughout the laboratory quake cycle.

The applicability of such a model needs to be further validated by the dataset of real-world earthquake events. In addition, similar to the previous models, it focused solely on the fault behavior at the local scale, which may be a shortcoming. Some researchers have noticed the potential interaction among seismic events, in the long-ranged and -delayed pattern [37,39,40,42-55]. Specifically, Raleigh et al. [36] observed that the seismic activities across ***California*** and the ***Middle America Trench*** are evolving in a somewhat synchronized manner. These two locations are also two of the most earthquake-prone areas worldwide with relatively complete catalogs [39,40]. In this paper, we developed and applied a new stochastic modeling algorithm to the seismic activities associated with these two regions to further examine the potential interaction between them. The two regions are separated by approximately 3038 kilometers.

3.3 Data Sources and Statistical Methodology

Although large earthquakes are rare, smaller seismic activities are somewhat pervasive at most space-time domains. The basic data source for these activities is provided by the seismic catalogs, which list the location, time, and size of earthquakes (and sometimes, the source mechanisms

also). In this paper, we use the USGS global seismic catalog [58] comprising (as of December 2016) over 935314 events of magnitude M≥2 since January 1970, with the rate and precision of recordings improving every year. We recognize that this catalog may be incomplete or have somewhat inconsistent temporal and spatial sampling, though it is the most comprehensive catalog that is publically available.

To probe into the potential causality pattern among these recorded earthquake events, we propose a new non-parametric computation of causality for quantized or symbolic data streams generated by ergodic stationary sources [38,59]. Ergodic processes are those for which statistical properties may be deduced correctly from a single, sufficiently long realization of the process. Stationary processes are those whose statistical properties remain unchanged over time. We assume that to a first approximation, seismic phenomenon are both ergodic and stationary, at least over the finite space-time scale of interest.

In contrast to state-of-the-art binary tests [60,61], our approach computes the degree of causal dependence between two data streams, without making any restrictive assumptions, linearity or otherwise. Additionally, without any a priori imposition of the specific dynamical structure, we infer explicit generative models of causal cross-dependence, which are then used for prediction. These explicit models are represented as generalized probabilistic automata, referred to crossed automata [59], and are shown to be sufficient in capturing a fairly general class of causal temporal dependence. The proposed algorithms are computationally efficient in the PAC [62] sense; i.e., we find good models of cross-dependence with high probability, with polynomial run-times and sample complexities.

3.4 Inferring Statistical Causality

Correlation measures are used to discern statistical relationships between observed variables in almost all branches of data-driven scientic inquiry. However, what we are really interested in is the existence of causal dependence. Statistical tests for causality, it turns out, are signicantly harder to construct; the diculty stemming from both philosophical hurdles in making precise the notion

of causality and the practical issue of obtaining an operational procedure from a philosophically sound denition. In particular, designing an ecient causality test that may be carried out in the absence of restrictive presuppositions on the underlying dynamical structure of the data at hand, is non-trivial. Nevertheless, the ability to computationally infer statistical *prima facie* evidence of causal dependence may yield a far more discriminative tool for data analysis compared to the calculation of simple correlations.

Nobel Laureate C. W. J. Granger, attempted to obtain a precise denition of causal inuence [38]. His denition proceeds with the following simple intuitive notion: Process A is a cause of process B, if process A has unique information that alters the probabilistic estimate of the immediate future of B, more so than prior values of B are indicative of the future of B. Not all notions of causal inuence are expressible in this manner, neither can all philosophical subtleties be adequately addressed. Grangers motivation was more pragmatic. He was primarily interested in obtaining a mathematically precise framework that leads to an eective or algorithmic solution - a concrete statistical test for causality. In this article, any reference to the term causality henceforth refers to this strict statistical denition of Granger-causality.

3.5 Spatio-Temporal Quantization

Before calculating the causal model and making predictions, we preprocess the USGS data within California and the Middle America trench regions, as shown in Figure. 3.1. We specify a magnitude quantization threshold, chosen to be 4 and 4.5 for California and the Middle America Trench, respectively. The threshold is chosen such that the catalog is considered complete because recordings fit the well-known Gutenberg-Richter law. This is the completeness criterion: below these thresholds we dont record all events and the Gutenberg-Richter power law is broken. This law suggests that earthquake magnitudes are distributed exponentially as $\log_{10} N(m) = a - \mathrm{b} \times m$ whereăN(m) is the number of earthquakes with magnitude higher or equal toăm,ăwhere băis a scaling parameter andăaăis a constant, given any region and time period [63]. For every week from the beginning of the dataset, we then assigned a “1” if there was at least one seismic event above

the tiles threshold, or "0" otherwise. For example, the sequence 000101 for California implies that at least one event at or above M 4.0 was recorded in both the 4^{th} and 6^{th} weeks since the beginning of the observation period. Therefore, our goal is therefore to test if one series of bits can help predict the second series of bits.

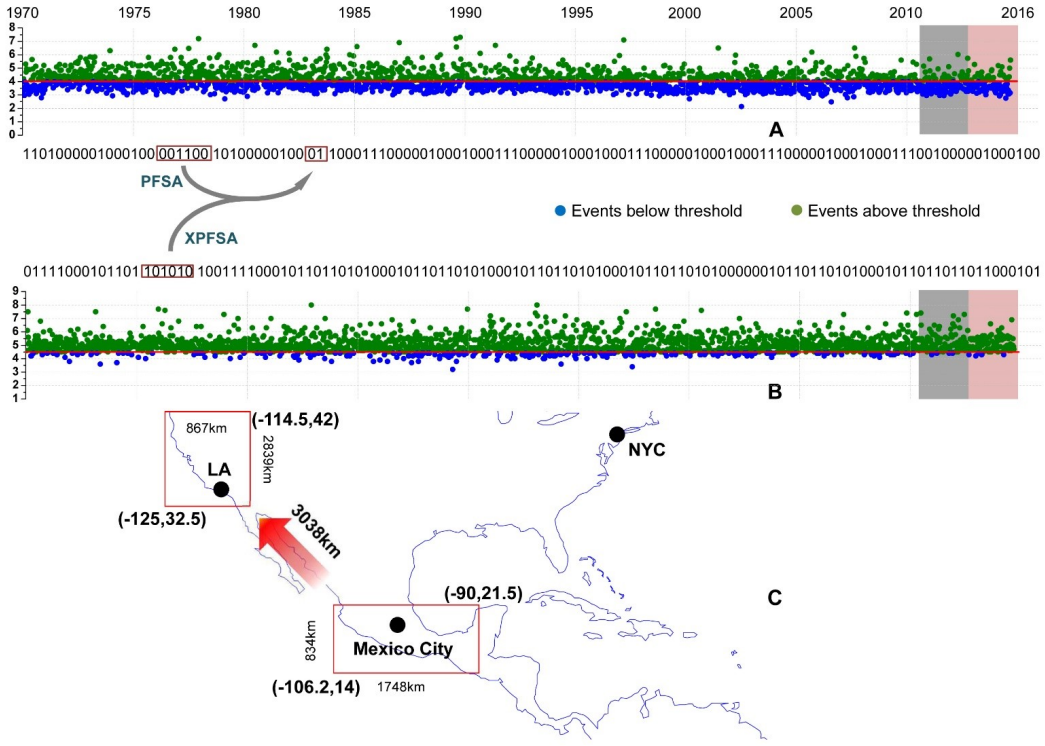


Figure 3.1: Exhaustive analysis of the earthquake catalog in California and the Middle American Trench (3038 km away) with different time delays. We quantized the time series data into 0s and 1s first (for California (Chart A) at a magnitude of 4 and the Middle America Trench (Chart B) at a magnitude of 4.5) and then infer XPFSA models with all-time delays.

Many other alternative quantization schemes are also possible, for example, we could use different thresholds, use different periods instead of a week, or use alternative regions instead of the specific rectangles chosen. We do not claim that the causal influence can only be revealed by this specific quantization scheme or these regions. However the specific quantization scheme chosen suffices to demonstration the existence of statistically-significant causality. As data quantity and quality improves in the future, better quantization methods and higher resolution quantization schemes will be possible and would undoubtedly produce even more precise causal relationship

calculations. For example, it may be possible to use days instead of weeks, smaller tiles, or use the earthquake magnitude instead of just a binary threshold of above or below a fixed threshold.

Once we obtain a series of ones and zeros for each region, we proceed to infer causal relationships between these two data streams, with various possible delays. Because the model determines to what extent activity in a one region predicts the activity in the second region in the future, we call such models cross models.

The extent of the delay between the cause and effect is unknown a priori. Therefore, we produce models for each possible delay value, ranging from one week to 600 weeks. We then choose the models as validated on a validation period whose AUC is above 0.5. The performance is then plotted on the test data.

3.6 Self Models, Cross Models, and Evaluation Metrics

There are serveral possible machine-learning techniques that can learn the relationship between a finite window of bits in a time series, and a future bit. We chose to use the probabilistic automata since they are proven to capture long-term relationships in stochastic processes [59,64]. There are two mathematical frameworks for modeling stochastic process. We call these two models Probabilistic Finite State Automata (PFSA) and Crossed Probabilistic Finite State Automata (XPFSA) [64]. Since we found out the long distance and long delay relationship in this paper validated by cross model, so our modeling method mainly focus on cross model and self model is also trained but only for comparison purpose.

Given two bit-sequences from different areas, we proceed to model the interdependencies between the series of bits representing the two regions using XPFSAs. Each state in an XPFSA contains the distribution of symbols of a dependent time series. Naturally, if stream A is not predictive of stream B, then the XPFSA G_{ab} will only be able to predict the average distribution of the symbols in B. However, if stream A contains information about the future of B, the corresponding model will be more specific in its predictions.

A cross model (XPFSA) can be viewed as a variation of a non-deterministic finite automaton.

It consists of four elements: a finite set of states, a transition arc, a driving symbol and a probability distribution within the state. If we take Figure. 3.2 (b) as an example, one can think of the cross model as trying to capture one stochastic process influencing (driving string) another (driven string). For example, you and another car are driving on a two lane road. The other car is always ahead of you and randomly changes lanes. Treat ***'0'*** as the changed lane and ***'1'*** as the kept same lane. No matter what the state was before, if the car ahead of you changed lanes, i.e., driving symbol is 0, the cross model state will always go to q_0, the probability distribution in state q_0, is 80% for 0 and 20% for 1. This means 'you' (the driven) will have a very **high** chance to change lane because of a direct causal influence, whether the the car ahead of you changed lane or not. If the car ahead of you did not change lanes, the state goes to q_1., given the state probability distribution is 20% for 1 and 80% for 0, 'you' will have a very **low** probability to change lanes. The cross model captures this stochastic causal relationship very well. For the self-model (PFSA), Figure. 3.2(a), instead of two cars are driving in two lanes, only you alone are driving. If you changed the lane last time, the self-model state goes to q_0 (20% for 0 and 80% for 1 on the transition arc). This means that because you changed the lanes last time, its highly possible that you will not change lanes this time. The PFSA captures a self-causal process: what happened in the past of the model itself has a causal influence on the model output in the future.

We perform two steps to train XPFSA which infers the strongly connected minimal realization (a detailed description of the algorithm can be found in [64]):

1. *Identification of ϵ-synchronizing string x_0*: Construct a cross-derivative heap using the observed traces with maximal length set as $\log\chi(1/\epsilon)$ where χ is the cardinality of the alphabet. For example, for a binary string (χ=2), and set as 0.15, then the traces maximal length is calculated as max($log_2(1/0.15)$)=3. The observed traces within this length are [null, 0, 1, 00, 01, 10, 11, 000, 001, 010, 011, 100, 101, 110, 111]. A cross-derivative heap is the set of probability distributions over cardinality for a subset of strings, which are the observed traces. For example, we calculate the traces tail 0/1 frequency ratio of all the observed traces within the length of 3 in the previous example: null,0,1,00,01,10,11,000,001,010,011,100,101,110,111.

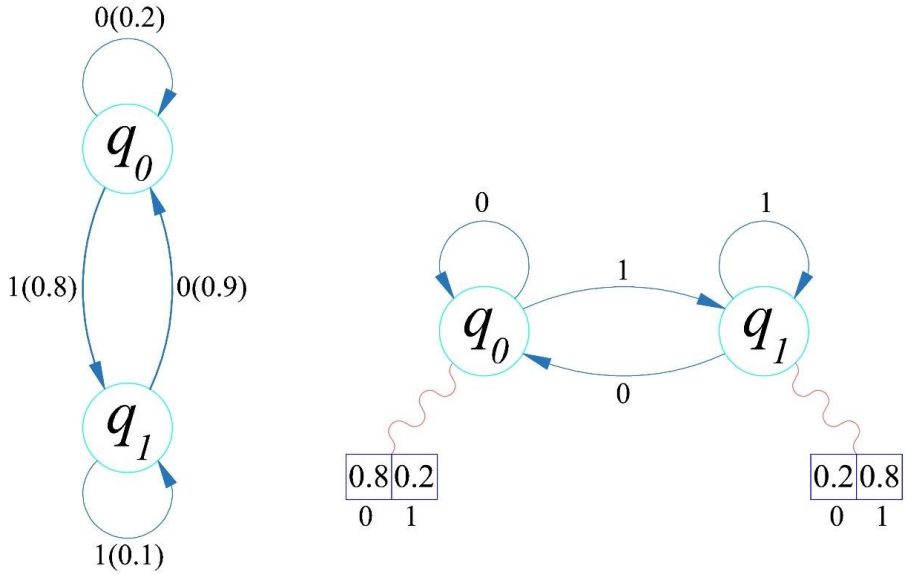


Figure 3.2: Illustrative PFSA and XPFSA models: there are two simple examples of PFSA and XPFSA, respectively. They all consist of four elements: state, transition arc, driving symbol and probability distribution. The difference is: for PFSA the probability distribution is over the arc, but for the cross model, the probability distribution is within the state. For these two state machines, the symbol '11' is their identical synchronization string [64]. This means that no matter which state is, after running '11' to trigger the transition, it will always end in state q1. For the PFSA, its a self-model, after run into substring '11' in the input string, next bit has 10% of probability to be as 1 and 90% as 0. Similar for cross model, because its a cross model, after run into substring '11' in the driving string, the corresponding driven strings next bit has 20% to be 0 and 80% to be 1.

Then, we use all tail frequency ratios to approximate probability distributions by treating the set of frequency ratios as a cross-derivative heap. We then identify a vertex of the convex hull for the heap, via any standard algorithm for computing the hull. Choose x0 as the -synchronizing string mapping to this vertex.

2. *Identification of the transition function*: We generate transition functions as follows:

   (a) Initialize the set Q as state q_0 ,set x_0 as q_0's string identifier and its tail frequency ratio as state probability distribution.

   (b) Then, we create a tree structure and set q_0 as the root, then compute for each trace

from that roots tail frequency ratio. If there exists a state q, the uniform norm of its and existing states probability distribution is smaller than ϵ, then we merge the state q into the existing state, if not, we define q as a new state.

(c) The process ends when a new state is no longer being defined.

(d) Then, if necessary, we ensure strong connectivity using Tarjans algorithm

Evaluating the performance of earthquake predictions is difficult. Anyone can make the trivial prediction that there will be no earthquake tomorrow and will have no false alarms, or zero false positives. Similarly, one can easily state that there will be an earthquake in the next 100 years and never miss an actual earthquake, thus having zero false negatives. However, both predictions are useless. The challenge is to have as few false alarms as possible, while simultaneously having as many true positives as possible.

One standard way of evaluating performance in such situations is the Receiver Operating Characteristic (ROC) curve, which was originally designed to rank radar operators in World War II. The ROC curve plots the probability of true positive rates as a function of false positive rates, as the decision threshold is varied. The area under the ROC curve, often referred to as the ROC area or Area Under Curve AUC, represents a convenient and statistically robust measure of classification performance. If the performace is worse than random, the ROC curve is a diagonal line, and the AUC is 0.5. In the perfect case where we generate 100% true positives while never generating false negatives, we have an AUC of 1.0. In practice, any nontrivial statistical causality will lie somewhere between 0.5 and 1.0.

We split 90% of the dataset for the training XPFSA model, 5% for the validation and 5% for the testing. We first computed the XPFSAs from the Middle America Trench quantized stream pointing to the California quantized stream. We tested all possible shifts, in one-week increments, for up to 600 weeks, equivalent to approximately 12 years. These models reveal to what degree the future seismic activity in California is predicted by activity in Middle America Trenchs long term and short-term pasts. Then, we picked only the XPFSA models whose individual prediction AUC

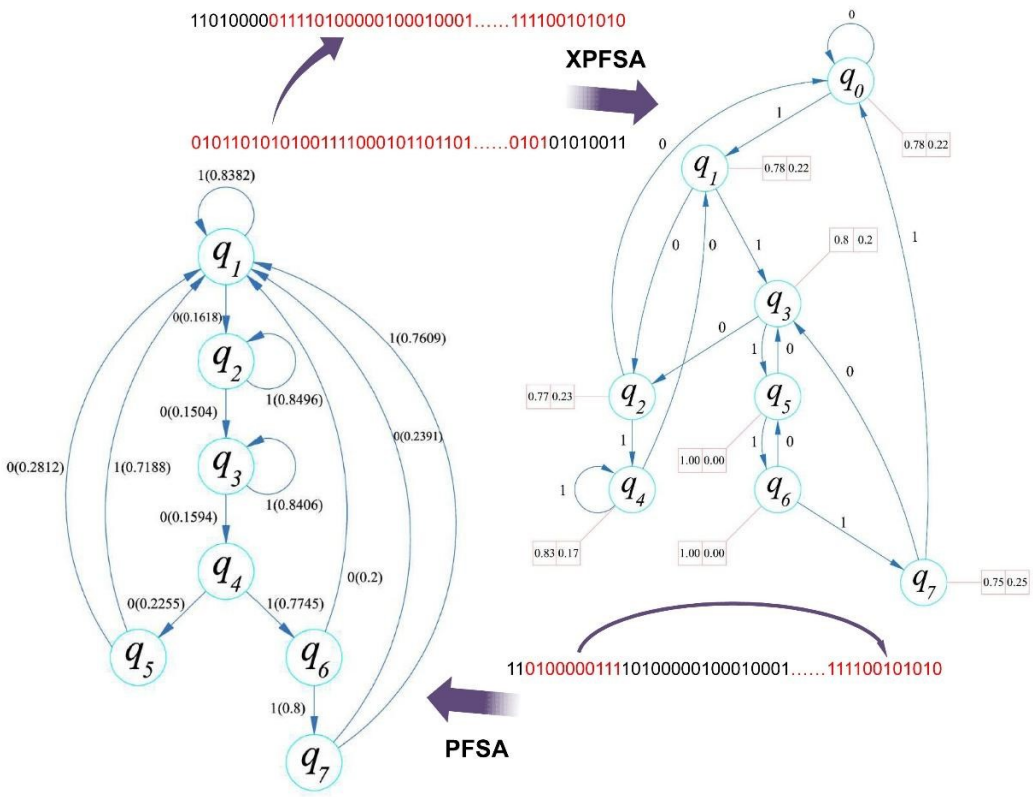


Figure 3.3: XPFSA (cross) model calculated from two quantized earthquake streams (California and the Middle America Trench) with time delays.

value was bigger than 0.5 in the validation set because these individual models are doing better than the random guesses in the validation set. Figure 4.1 is one example of the trained XPFSA.

The next step in obtaining future predictions is fusing the predictions from the individual models. Here, we simply average the individual predictions, though more sophisticated methods could be explored in the future.

Is this framework powerful enough to undercover the hidden causal seismic dynamics? We provide an affirmative answer to this question, at least with respect to the quantization schemes we adopt, by showing that the prediction AUC value in the test dataset is high and unlikely to be pure chance. In other words, the predictive causality is statistically non-trivial.

If one examines the AUC value combined from different time delays, it is apparent that neither the shortest time delay combined model (combining models in the range of 0-100 weeks delays), nor the longest delay (comprising models with 500-600 week delays) provide the most informative

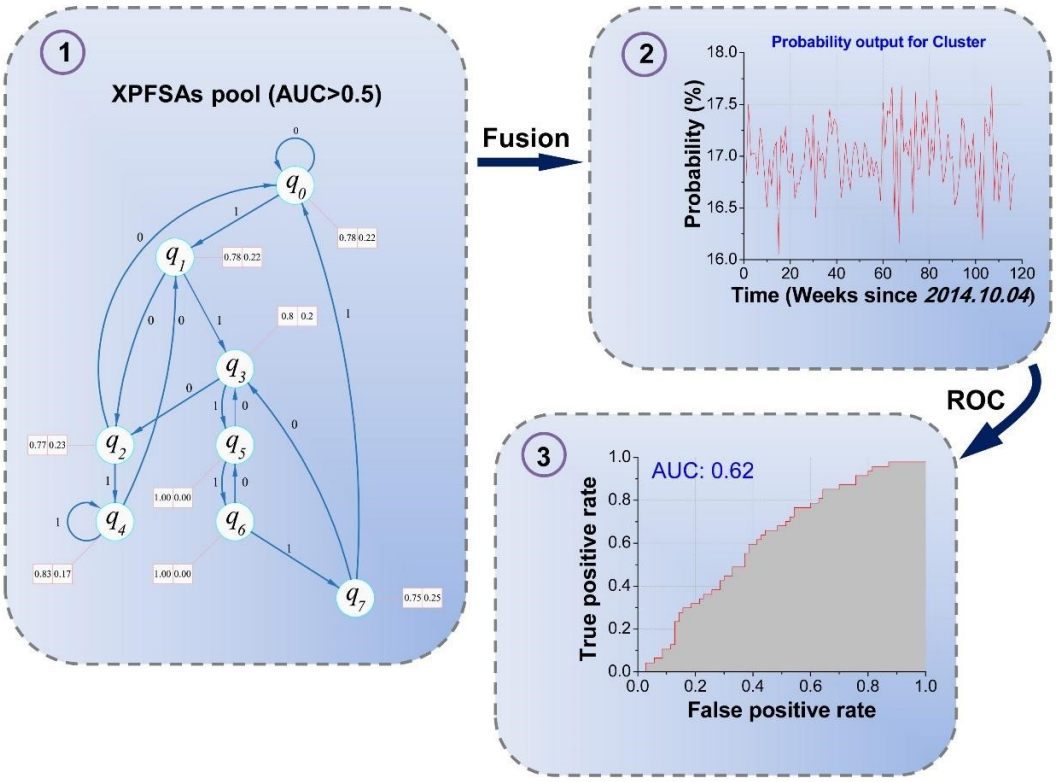


Figure 3.4: After the individual XPFSA, whose AUC in validation is larger than 0.5, is chosen, we averaged their outputs and fuse them into one vector prediction. Finally, we calculated the AUC of the fused vector predictions against the test dataset.

prediction information to the targeted area. However, the combined prediction of models with delays in the range of 250 to 350 weeks turns out to be the one with the highest AUC value (Figure. 3.4). This result suggests that, from a data-driven perspective, there exists on average,a 6-years delay causation effect from the Middle America Trench to California.

3.7 Control Experiment and Sensitivity Analysis

To show that the AUC value of the out-of-sample data is statistically non-trivial, we first shuffled the time series for both regions and recomputed the predictions. Specifically, we shuffled the data in large 3-month sets. Such coarse shuffling serves to destroy any long-term causality relationships while retaining the short-term characteristics of the data. As expected, the average AUC value of the shuffled data was close to 0.5, implying no causality.

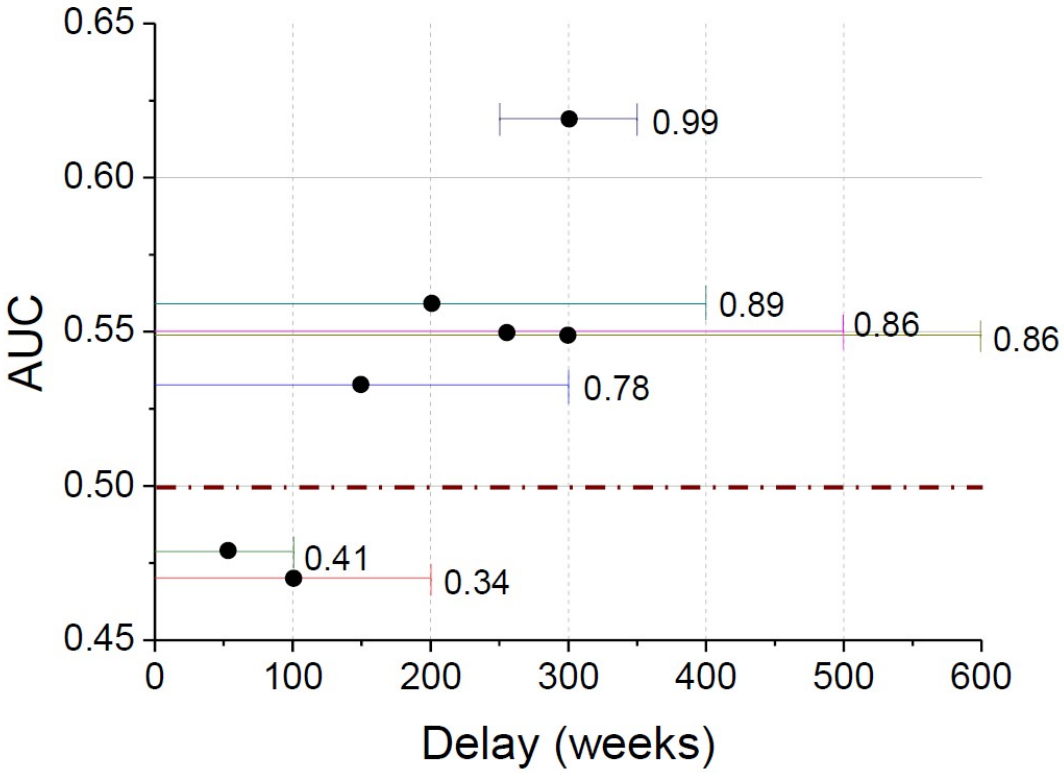


Figure 3.5: Comparison of models with various time delays. Different tie delays and durations result in different AUC. The duration (in weeks) is represented by a horizontal line and the statistical confidence is represented as a number to the right of the line. The highest AUC and confidence level is achieved by adding up delay from 250 weeks to 350 weeks. The default AUC of 0.5 is highlighted as a dotted line for reference.

Additionally, we computed the confidence level of the AUC value obtained by the XPFSA models. After the AUC values of many randomly coarse-shuffled time series were calculated, we integrated their area from - to where the original non-shuffled AUC value lies. This area corresponds to the statistical confidence level. The AUC value reported of 0.61 corresponded to a confidence level of 99.11%, as shown in Figure. 3.5 (highest deep blue line). In other words, the AUC value of 0.6191 has a less than 1% (1-99.11%) chance of being a coincidence.

As aforementioned, we also tested the result robustness by choosing different quantization scheme for California and Middle America Trench. The AUC values for these alternative quantization schemes all exhibit high confidence levels which also indicates a strong causal influence from Middle America Trench pointing to California. For example, after quantizing Middle Amer-

ica Trench at magnitude 4.6 and California 4.1, combining model delay from 250 weeks to 350 weeks reported AUC 0.62 and confidence level 99.2%. Quantized Middle America Trench at 4.7 and California 4.2, fused delay model also from 250 weeks to 350 weeks yielded AUC 0.58 and confidence level 95.28%.

In addition, as shown in Table 3.2, aside from the XPFSA, we also calculated the PFSA for comparison purposes. When it comes to smaller magnitudes (first two rows), the self-model and cross-model both shows high AUC. However, the cross-model performs much better when it comes to larger magnitude earthquakes.

Table 3.2: AUC Under Different Quantitation Thresholds

California, Middle America Trench(Magnitude)	**Self model (PFSA)(AUC)**	**Cross model (XPFSA)(AUC)**
CA: 4, MAT: 4.5	0.65	0.62
CA: 4.1, MAT: 4.6	0.65	0.62
CA: 4.2, MAT: 4.7	0.41	0.58

Besides different quantitation thresholds setups, the inverse causal direction is also tested for California to see whether is has the same causal influence on the Middle American Trench. With the same setup, the AUC is low as 0.51 which means the causal relationship is unidirectional.

To study the smaller area within and around the Middle American Trench, zoom in sensitivity analysis inside the Middle American Trench area is also implemented. We set the size of the box to be 1/4 of the original size of the Middle American Trench box, then we moved the center of the zoom in box incrementally in both vertical and horizontal directions to scan the entire original Middle American Trench. In Figure.2.6, on the left, we showed on the left how the boxes are moved and. On the right, we showed the associated AUC heat map. One can see a high AUC area clustered in the upper left corner (circled in red) in the heat map. The highest zoom in the boxs AUC is 0.67, which is even higher than the entire MAT box (0.62). This means that within Middle American Trench, strong influence sections exist. The coastline is also overlaid in the Figure. 3.6's left plot. On the far upper right and far lower left corners, the AUC values are close to 0.5, because the data in the open sea area are relatively sparse.

Since we captured the causal relationship starting from the middle America Trench to Califor-

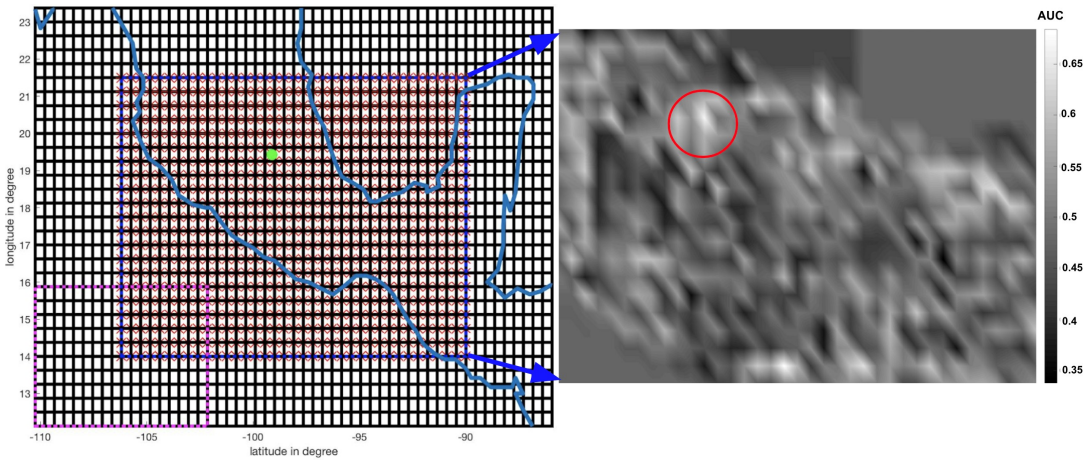


Figure 3.6: Heat maps for zoom in sensitivity analysis. We zoomed in and set the size of box to be 1/4 of the original size of the Middle American Trench box. Then, the center (red asterisk) of the zoom in box was moved incrementally in vertical and horizontal directions to scan the entire original Middle American Trench box (navy blue rectangle box in the middle of the Figure 2.6 left plot). The pink dotted box is one example of the zoom in box. The red asterisks represent all the centers of all moving the boxes. On the right is the heat map of AUC value of all moving boxes. Mexico City is plotted as a green point and the adjacent coastline is also overlaid in a bold blue line as a geographical location reference.

nia by using past earthquakes data, there is nothing holding us back from examining the driving sources from other areas. We used the same approach to search for other areas inflicted by earthquakes that have high causal relationships to California.

From the plot below(Figure. 3.7), the yellow indicates the Middle American Trench, which is researched in the previous chapter. The other high causal driving earthquakes are plotted using red dots. The surprising discovery was how **far** away these driving earthquakes could be, which suggested a long distance relationship between global earthquake activities.

Furthermore, we can use this approach to map out all the causal relationships globally. Detailed information can be found in Appendix A.

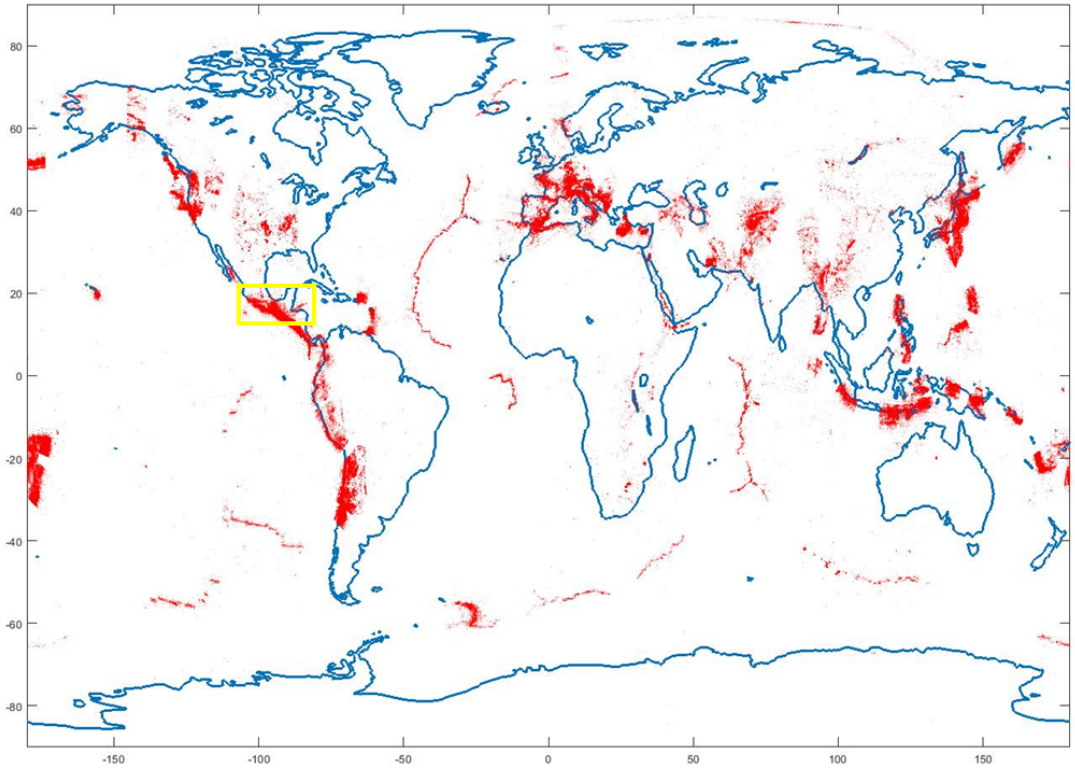


Figure 3.7: All clusters to LA: Ran k-means for 20 times, overlay all the clusters(from 20 k-means) if its test AUC is higher than 0.6. For each point, certain transparency is assigned for each point. The driving earthquakes could be very **far away**, which suggested a long distance relationship of global earthquake activities.

Chapter 4: A knowledge-Free Approach for Brain Activity Classification From Single Streams of Data

4.1 Introduction

Although Brain Computer Interface (BCI) techniques have had much progress in accuracy, speed, and usability in recent years, they are still encountered to have two significant limitations. i) The need of extensive background information on the type of signal being used to apply appropriate data processing and feature extraction techniques[88,89,90]. ii) Limited access to sources of signals i.e. neural activity[refs]. Various types of neural data are being used in BCI applications, including Electroencephalogram (EEG) [70], Electrocorticogram (ECoG) [71], Local Field Potentials (LFPs) [72], and Single Unit Activities (SUAs) [73], [74]. These are the most commonly used signals for BCI applications. Each of the brain imaging techniques have their own advantages and limitations. Regardless of these advantages and limitations, they require a specific approach to pre-processing data (filtering, decomposition of signals (ICA), dimension reduction (PCA), spike sorting, etc.) and extracting the task relevant features. The appropriate selection of processing methods depends on a strong knowledge of the studied system as well as finding the most relevant features to the task being decoded. Despite taking the most recent findings into account, our understanding of how the brain works is very limited. Therefore, it is not surprising that a single approach or at least a convergence to an ideal technique on how to process any of these signal groups exists. In fact, processing methods for each of them are open areas in research even today. Although having the maximum amount of information is always desirable, this amount is always limited by factors such as practicality, invasiveness, noise and artifacts, and limitations in accessing different parts of the brain. Although having access to as much information as possible is very important in both research and commercially available BCI products, it is even harder to retrieve

information in the second group due to the limitations that practicality and cost would impose. It is important to note that even though the input data is much richer than the commercial products in clinical studies, it is still considered very limited compared to the complexity of the studied system, which is again a limiting factor in the progress of both BCI and neuroscience fields. Here, we are proposing data-smashing for BCI, a method that overcomes both above-mentioned sets of limitations and offers much more. This chapter is organized as follows. Section 2 and 3 reviewed the literature and explain the motivation. The methods and significance of our approach are explained in Section 4. Section 5 summarizes the results. Final conclusions are provided in Section 6.

4.2 Related Work

In common approaches to solving machine learning classification problems used today, typically hundreds, thousands of examples are fed into the model for training. However, without prior knowledge, biological systems can learn from only a few examples[78,79,80,81]. Li et al[82] explored unsupervised one-shot learning of object categories in visual problems using a Bayesian approach. In these studies, a probability density function on the parameters represent prior knowledge. By updating the prior distribution in the light of one or more observations, the new category is obtained. However, this approach[83] requires building up the prior knowledge base of generic knowledge which may be obtained from previously learnt models of unrelated categories. In a more advanced deep learning approach in [84], the authors also showed how to use an extensive amount of knowledge about the visual world available in natural language to classify unseen objects. However, this also requires extensive training data to calculate the prior knowledge distribution. The same attempt is also made in discovering the new drugs. Han et al,[85] proposed a similar pipeline in learning the behavior of compounds in a new molecular scaffold, given only a few data points from the new scaffold. For the same nature of the dataset of this paper, Abbas, Alessio et al [86,87] applied one-shot classification on the EEG dataset to classify diseases such as seizures. Nevertheless, their algorithm requires EEG time series from all electrodes. However, these data are always limited by factors like practicality, invasiveness, noise and artifacts, and lim-

ited access to different parts of the brain. These drawbacks drive us to explore the prior-knowledge free method, which only requires a single source (one electrode) of brain activity to perform the classification.

4.3 Method and Significance

Data smashing is an algorithm, which, when given a pair of signals, calculates the statistical 'distance (dissimilarity) between their generator functions without the need of prior knowledge about the signals or the system that has generated them. These dissimilarity values are further used to classify the input signals which are now represented in the n-dimensional 'statistical distance space. This method can provide high performance even in the presence of the limitations mentioned earlier.

First, this method does not need any expert supervision, prior information about the system, or the nature of the data being fed to it.

Second, it is very effective on decoding desired states from, as little as, only one stream of data (e.g. one EEG or ECoG channel or a single SUA unit) where all other state of the art methods fail to operate accurately with this limitation. We tested this method to classify similar tasks (moving of the index finger versus moving of the thumb) from the data from a single ECoG electrode.

In addition, since no bias exists, this processing method can reveal new aspects of the nervous system in an unprecedented way and without limiting them by constraints set by prior assumptions.

We believe that it can revolutionize the commercial BCI products and potentially improve the practicality of brain activity recording devices (e.g. NeuroSky MindWave headsets) and similar products, as well as improving our understanding of how the brain functions and the field of neuroscience overall. We have provided a short introduction to the data-smashing method in Chapter1 Section 1.8.

4.4 Experiments and Results

Here, we have provided results from running data-smashing on two BCI applications with fundamentally different recording modalities.

4.4.1 *Decoding Finger Flexion from Single ECoG Signals in Humans*

In this experiment, users were asked to move their fingers (one at a time) while their ECoG signals and finger movements were being recorded simultaneously. The dataset is fully described in [75]. The goal here was to be able to distinguish movements of index versus thumb finger movements.

Figure. **??** shows the probabilistic distance of one channel of the ECoG recording for the two different classes in the reduced dimension space as well as the convex-hulls that they lie within. Even though there is more than one way to reduce dimensionality and visualize data in the lower dimensional spaces, they all lead to similar results. Here, we used PCA specifically and choose the first three principal components. The algorithm could differentiate the two classes in the probabilistic space even in the absence of any prior knowledge on signals or any expert modifications on the code specific to the nature of the data.

Our algorithm could classify the two classes (thumb movement or index movement) from only one ECoG channel with the accuracy of %91.17. The average accuracy for the best five channels were %88.82 and the histogram for the accuracy among all the electrodes is presented in Figure. 4.2. Original data labels were used to interpret the unsupervised classification results made by the algorithm.

These results are very surprising since this method was performed without the need for signal specific pre-filtering, feature extractions, or any domain knowledge. This algorithm is also generalizable and we have tested it on other BCI signals, such as EEG [76], and Single Unit Activities (Neural Spikings). Without using any prior knowledge, it shows promising performance compared to other state of the art techniques[71,91].

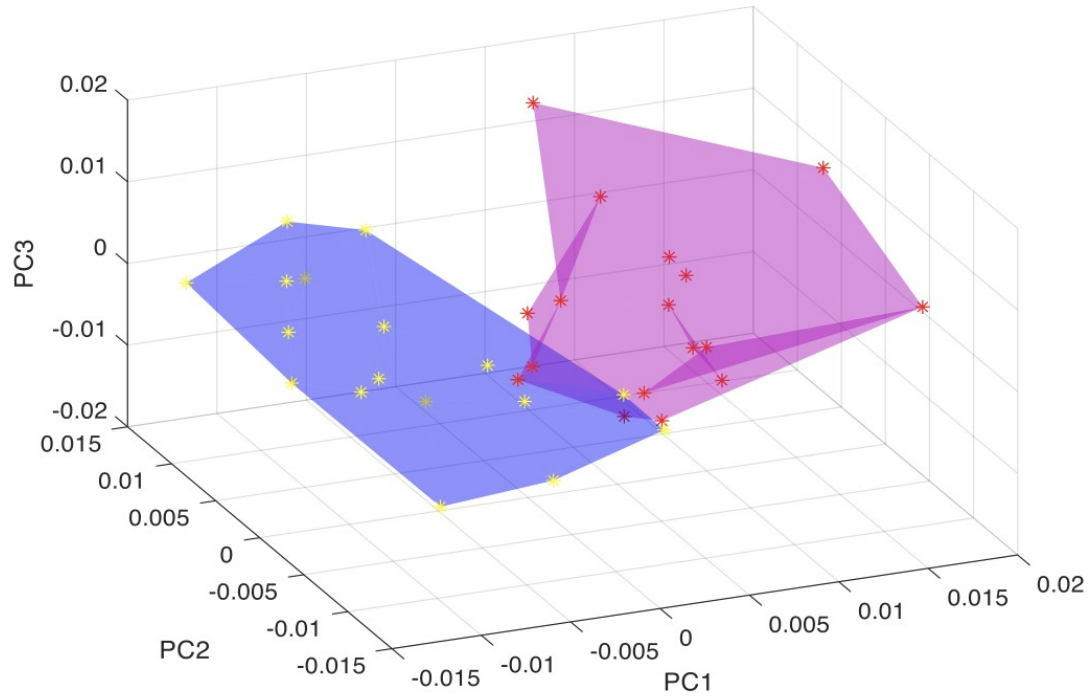


Figure 4.1: The similarity distance representation in the PCA space. It is clustered into two groups: a red convex hull group with red dots representing the samples, and a blue convex hull group with yellow dots representing the samples.

Thanks to its revolutionary approach to the signals and without prior filtering, signal conditioning, and feature extraction, data-smashing is efficient. It takes less than 0.25 seconds to smash a pair of two ECoG signals (with each having 5347 samples) with a commercially available computer (exact specifications of the computer system are available in appendix A). Moreover, since the data-smashing is a completely parallelable algorithm, its speed for smashing multiple pairs can be greatly increased with parallel computation platforms such as CUDA [77] and cloud computing. Therefore, it is a very good candidate for real-time BCI applications.

4.4.2 *Detect **Single**-Cell Activity Level in Mice Brain*

The second application we applied this methodology on is mice brain activity data from the *Allen Brain Observatory*[122]. This open database presents the first standardized in vivo calcium imaging of physiological activity in the mouse visual cortex, featuring representations of visually evoked neuronal responses from GCaMP6-expressing neurons in selected cortical layers, visual

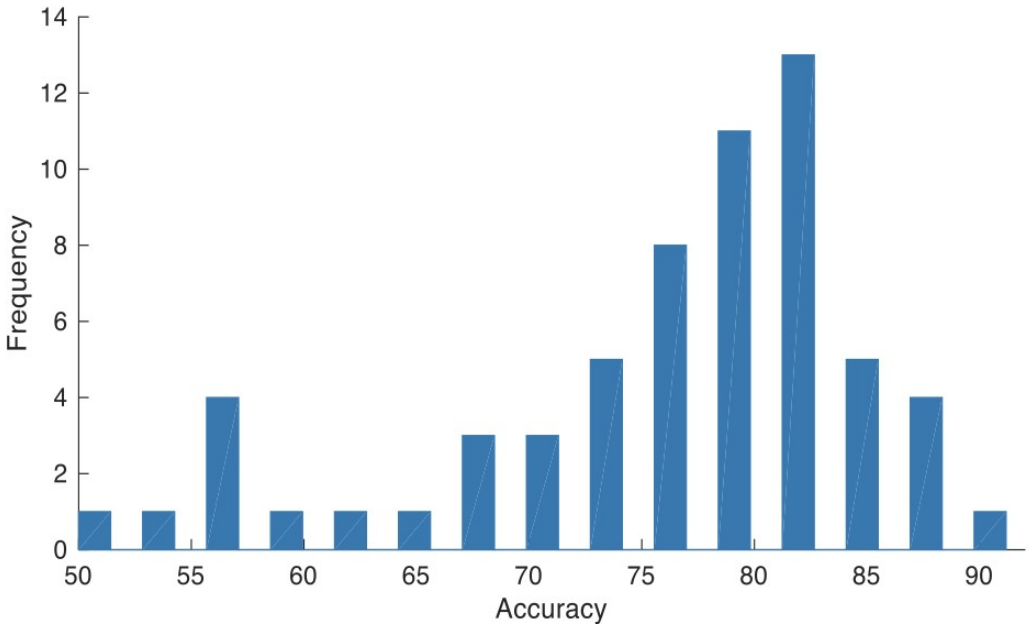


Figure 4.2: The histogram of accuracy for all ECoG channels.

areas. Given that the physiological response of the mice under visual stimuli is unknown and true label is not available, we want to data smash the signals and observe what pattern would emerge after embedding.

The specific dataset we used in this paper are experiments 511510855, 511510670, 511510650, 511507650, 511509529. Within each experiment, different stimuli are being presented to mice at different periods of time, including drifting gratings, static gratings, natural scenes, natural movies, etc. Detailed information can be found here [122]. We focus on the period where grating with different orientations is visually presented to mice.

The idea is to apply data smashing on all time series of cell recordings within one experiment to examine whether the cell embedding in the space has a certain pattern along the manifold that reacts to any physiological state of the mice.

For all five experiments, data smashing results provide us with a distinctive pattern for different cell reactions.

After data smashing, all five experiments cell similarity relationships are embedded and plotted in the first two principal components spaces. The shape of all five experiments exhibited similar

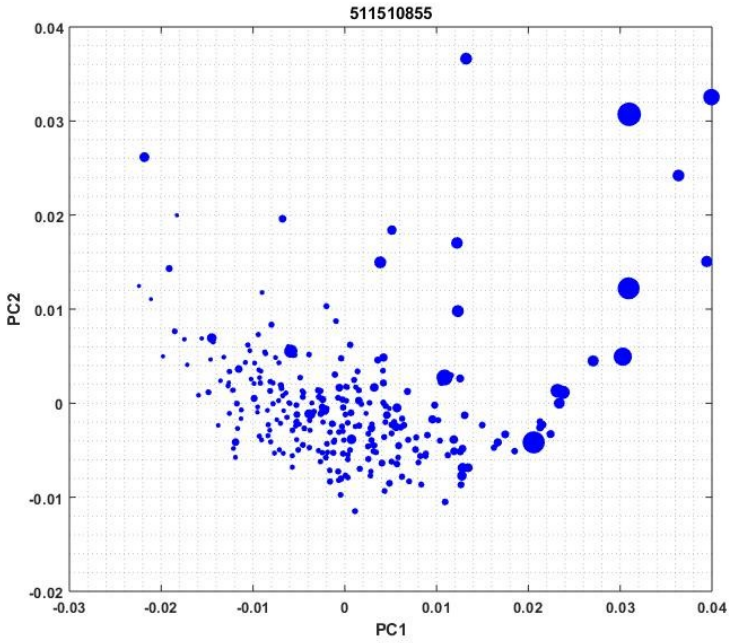


Figure 4.3: Activity level proportional to size of the dots in data smashing space(first two principal components) of experiment 511510855

patterns and a U-shape manifold was observed. After a few trial and error attempts, we realized that the pattern is correlated with the cell activity level. To get a direct visual sense of the cell behavior along the manifold, we plotted the dots radius proportional to each cells activity level. We defined the percentage of spiking time in each neuron as the activity level. There are many spike inference algorithms available. Here, we simply calculated the first derivative of dF/F traces, and defined everything above mean+3*SD as spike events.From Figure. 4.3 to Figure. 4.7, the radius of the dot on the right gets larger when compared to the left, where the radius is positively correlated with each cell activity level.

The mechanism behind this pattern of cell behavior remains unknown. It is not clear what the difference between the cell clusters on the right and clusters on the left are. However, the pattern itself is self-evident and the method enable this discovery required zero prior knowledge or expertise in this specific field.

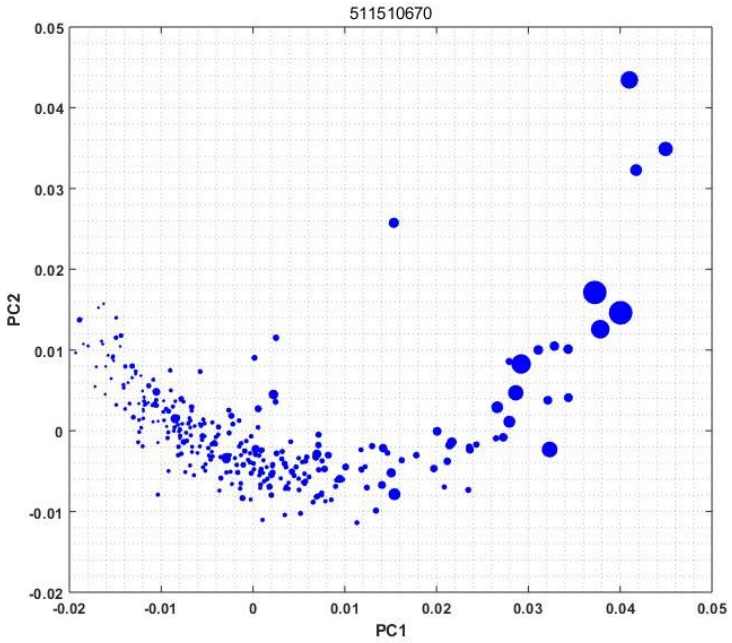


Figure 4.4: Activity level proportional to size of the dots in data smashing space(first two principal components) of experiment 511510670

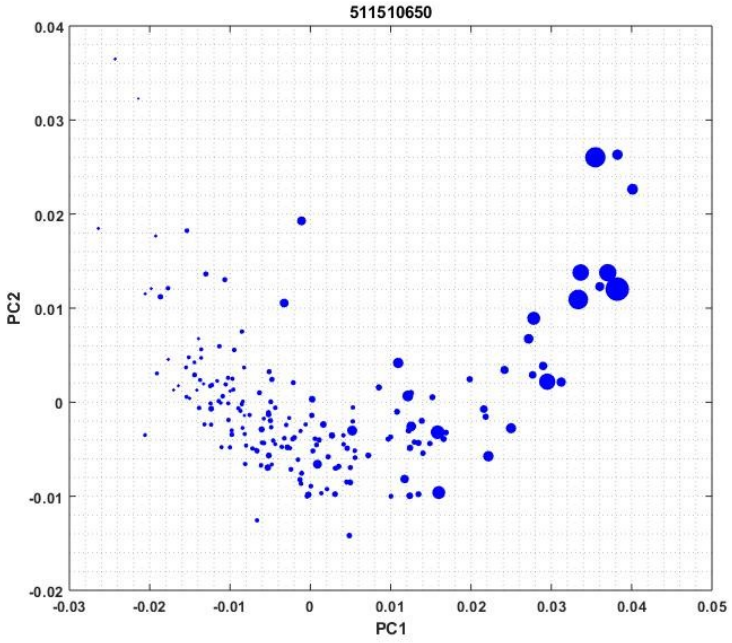


Figure 4.5: Activity level proportional to size of the dots in data smashing space(first two principal components) of experiment 511510650

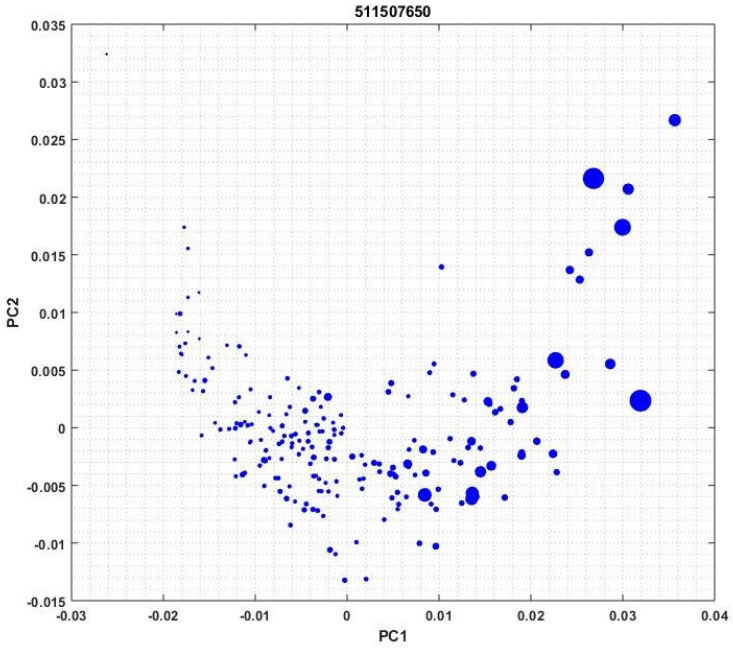


Figure 4.6: Activity level proportional to size of the dots in data smashing space(first two principal components) of experiment 511507650

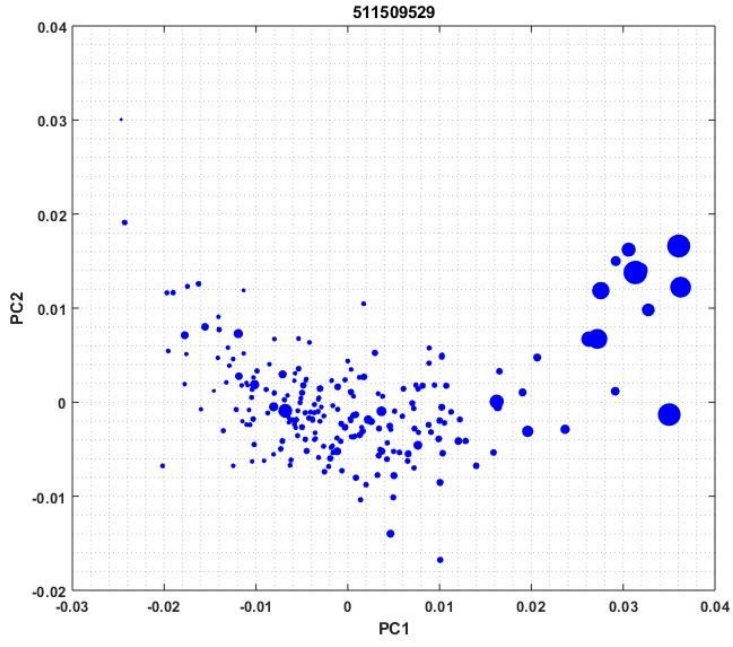


Figure 4.7: Activity level proportional to size of the dots in data smashing space(first two principal components) of experiment 511509529

Chapter 5: Non-Parametric Distribution-Free Metric for Dataset Predictivity Estimation

5.1 Introduction

The accelerating abundance of data is fueling the potential of computational sciences across many disciplines and industries. However, our ability to predict which datasets are likely to be useful is not keeping pace. The data varies from sensor data streams from of everyday electronic devices to recorded cosmic radiation, as well as social constructs such as stock market prices.

With the prevalence of the data[115], various predictive methods such as machine learning models are being developed and deployed by doing prediction and classification[116]. As the machine learning algorithms themselves become fairly standardized, a key challenge for data scientists is to determine which dataset is most useful for the task at hand. Typically, practitioners will simply experiment with various algorithms to see what works and what does not, based on their personal experience.

While there has been relatively abundant research on developing, testing, and comparing performance machine learning algorithms on standard datasets, relatively little has been done to determine which datasets are inherently more amenable to machine learning. Here we ask whether it is possible to determine ***a priori*** how likely a given dataset is to be predictive under any machine learning algorithm. While answering this question is impossible, we aim to find a metric that would correlate with future performance of established standard machine learning methods. We specifically focus on time series data.

For time series data, a simple way to do this is to calculated the Shannon entropy[118]. Shannon entropy is a metric for how chaotic the system is. Before feeding the data into any machine learning system,the calculated Shannon entropy could be one simple index for how predictable the

dataset will be. However Shannon entropy sometimes oversimplifies the representation of the data. For example, for time series data, when calculating the Shannon entropy, it is possible to overlook the long-term temporal patterns hidden in the data stream.

Here, we introduce a new non-parametric method to calculate a quantized data streams chaotic level, a value which can be treated as a dataset predictivity metric. Our method does not require any prior knowledge of the data nor of the system that generated the data. It provides a metric of predictability that allows us to adjust our expectation of the prediction outcome to help us prioritize and allocate our resources to solvable systems, ultimately increasing the overall efficiency of limited computational resources. Importantly, the predictive measure we propose can be calculated in linear time, and is, thus, faster and cheaper to calculate compared to running a full blown machine learning session.

This paper is organized as follows. Section 5.2 and 5.3 review the literature and explain our motivation. The overview and discussion of our approach is explained in Section 5.4. Section 5.5 summarizes the experimental results. The discussion and final conclusions are provided in Section 5.6 and 5.7.

5.2 Related Work

Several attempts have been made to conduct the predictability analysis in specific fields, from stock market returns to absence seizures, from predicting human mobility to network traffic and atmosphere [107]. Song, Qu *et al* [100][101][102] utilized entropy to approach the limits of predictability in human mobility. Ding *et al* [102] tried using entropy to study the degree of radio spectrum state predictability.Maasoumi *et al* [95] examined the predictability of stock market returns by using an entropy metric. They found that entropy can detect nonlinear dependence within series return. Their results indicate that some of the inference is sensitive to the period of analysis and other factors. Also, they used second-order Gaussian kernel for the kernel function, whereas the method in this paper is distribution-free.

The approximate entropy measure is introduced in [96], as a rate of entropy for an approxi-

mating Markov chain to a process. Richman *et al* [97] claimed a statistically superior method than approximate entropy and a related complexity measure called *sample entropy*. In [98], permutation entropy is developed as a tool to predict the absence seizures of genetic absence epilepsy rats. Li *et al*[98] discovered that permutation entropy can track transient dynamics before absence seizures. They detected 169 out of 314 seizures, at a rate of 53.8%, which significantly higher than the rate that is achievable using sample entropy, at 21%. However, approximate, sample and permutation entropy methods are parametric.

Molgedey *et al*[99] also explored the predictability of financial time series using local order and conditional entropy. What they discovered is that even when the financial time series is nearly random, special local situations might exists, where locality is present and the predictability is higher than average. In this paper, we focus more on measuring the given dataset predictability overall, instead of using a moving window.

Krumme *et al*[106] analyzed predictability of consumer visitation patterns. In the paper, the authors use an estimate of sequence-dependent (SD) entropy to measure, and a set of Markov Chain models to predict the location of shoppers. The Lempel-Ziv[104][105] algorithm was applied to estimate SD entropy. Bubble entropy was developed as an advancement of permutation entropy with the aim of reducing the significance of parameter selection. However, both papers are not parameter-free.

5.3 Motivation

Currently, many data-science researchers focus on algorithm choice and optimization. Benchmarks are conducted by running standard datasets through new algorithms and then comparing test accuracies. A higher accuracy typically indicates that an algorithm performs better. However, high test accuracy on standard datasets does not necessarily mean the model also performs well, or generalizes, on other datasets. The generalization capacity is a measure of how accurately an algorithm can predict outcome values for previously unseen data [117]. Simply coming up with an algorithm that performs well on a large dataset, such as MNIST or ImageNet, is no guarantee

for success on datasets from other sources. Here, we propose to make the dataset the focal point, rather than the algorithm.

The question we aim to answer is: Regardless of the machine learning model used, is there a way to assess the dataset predictivity before feeding the data into an algorithm? We restrict our claim to a conventional algorithm; however, we hope our approach generalizes to future algorithms as well, if the predictivity is a property of the dataset and not the algorithm. For example, no algorithm can predict the next bit in a series of coin tosses, no matter what algorithm is used. Therefore, the predictivity of the dataset should be zero, indicating that any attempt to predict is futile.

In this paper, we focus on quantized time series data. For quantized time series data, a naive way to measure the chaotic level of the system is to calculate its Shannon entropy. Shannon entropy estimate the average rate of information that is produced by a stochastic source of data[113]. In the other words, this is the unpredictability of the state, or equivalently, its average information content.

However, Shannon entropy is the average rate of information. The definition restrains its ability to capture the temporal memory of time series data. This urges us to look for a better metric that can also be easily calculated and one that could capture temporal relationships at the same time.

5.4 Data Smashing Metric for Data Predictivity

Earlier, we introduced an unsupervised learning method called data smashing[13]. Data smashing is used to measure causal similarity between series of quantized sequential observations. As a feature-less model-free classification method, data smashing does not require training, or expert tuned heuristics. It also provides flexibility on inputs data: equal length of time series is not required, and mismatch in phase and missing data can be tolerated. Once the similarity between the two time series is calculated by data smashing, one can use the resulting metric to cluster and compare datasets.

A very simplified process of data smashing can be found in Chapter 1 ,Section 1.8:

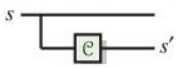
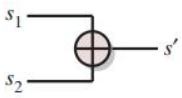
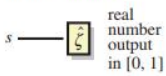
stream operation	algorithmic procedure (pseudocode)
independent stream copy[a]  generate an independent sample path from the same hidden stochastic source	(1) generate stream ω_0 from FWN (2) read current symbol σ_1 from s_1, and σ_2 from ω_0 (3) if $\sigma_1 = \sigma_2$, then write σ_1 to output s' (4) read next symbol and go to step 1 *this operation is required internally in stream inversion*
stream inversion[a]  generate sample path from inverse model of hidden source	(1) generate $\lvert\Sigma\rvert - 1$ independent copies of s_1: $s_1, \cdots, s_{\lvert\Sigma\rvert-1}$ (2) read current symbols σ_i from s_i $(i = 1, \cdots, \lvert\Sigma\rvert - 1)$ (3) if $\sigma_i \neq \sigma_j$ for all distinct i, j, then write $\Sigma \setminus \bigcup_{i=1}^{\lvert\Sigma\rvert-1} \sigma_i$ to output s' (4) read next symbol and go to step 1
stream summation[a]  generating sample path from sum of hidden sources	(1) read current symbols σ_i from s_i $(i = 1, 2)$ (2) if $\sigma_1 = \sigma_2$, then write to output s' (3) read next symbol and go to step 1
deviation from FWN[b]  estimating the deviation of a symbolic stream from FWN (symbolic derivatives (electronic supplementary material, Definition S-9) in the electronic supplementary material, Section S-B, formalize $\phi^s(\cdot)$. If s is generated by a FWN process, then $\phi^s(x) \to \mathcal{U}_\Sigma$ for any $x \in P\Sigma^\star$, and hence $\hat{\zeta}(s, \ell) \to 0$)	$\hat{\zeta}(s, \ell) = \frac{\lvert\Sigma\rvert - 1}{\lvert\Sigma\rvert} \sum_{x:\lvert x\rvert \leqq \ell} \frac{\lVert\phi^s(x) - \mathcal{U}_\Sigma\rVert_\infty}{\lvert\Sigma\rvert^{2\lvert x\rvert}}$, where — $\lvert\Sigma\rvert$ is alphabet size, $\lvert x\rvert$ is the length of string x — ℓ is the maximum length of strings up to which the sum is evaluated. For a given $\epsilon^\star$, we choose $\ell = \ln(1/\epsilon^\star)/\ln(\lvert\Sigma\rvert)$ (see the electronic supplementary material, Proposition SI-15) — $\mathcal{U}_\Sigma$: uniform probability vector of length $\lvert\Sigma\rvert$ — for $\sigma_i \in \Sigma$, $\phi^s(x)\rvert_i = \frac{\text{number of occurrences of } x\sigma_i \text{ in string } s}{\text{number of occurrences of } x \text{ in string } s}$

Figure 5.1: A detailed Algorithms for Data Smashing stream operations. Reproduced from [13].

If the inverted copy of one stream can annihilate the statistical information contained in the other, then we can claim that two sets of time series have the same underlying generative process without explicitly knowing or constructing the models themselves. In [110], this property is called information annihilation. In other words, if the input data stream s1 is statistically similar to a random stream s2 (which is known to be unpredictable by construction), then we conclude that the original data stream s1 is also unpredictable.

As shown in Figure 5.1(d), ***e12*** is the smashing distance between ***s1*** and ***s2***. The key aim of this paper is to ***arbitrarily set one of the streams as flat white noise***, as shown in Figure. 5.2.

If the given streams entropy level is high, close to a coin toss, the smashing distance would be close to zero because their underlying generative mechanism are similar. Conversely, the smashing

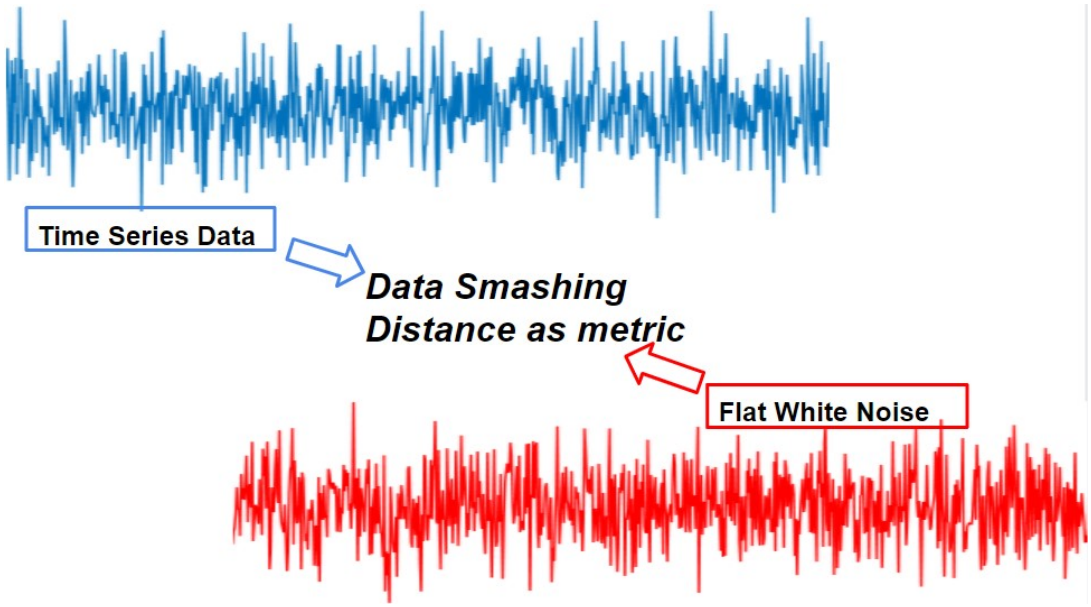


Figure 5.2: Arbitrarily set one of the streams as flat white noise, then perform data smashing with the original dataset.

distance would be much higher if the given time series data structure is close to deterministic because of the obvious difference between their underlying models.

One major advantage of using our approach compared is that when one smashes two time series to calculate their distance, the length of the data stream is finite. However, by using the proposed predictivity index we control the generation of the flat white noise, so we can create **FWN** (flat white noise) of any length needed. Just as with any machine learning method, having more data provides a higher chance of accurate convergence.

5.5 Implementation and Experimental Evaluation

To prove our metric is a better measure than simple Shannon entropy, we compared the correlation between validation accuracy and data smashing distance, to the correlation between validation accuracy and Shannon entropy. In other words, correlation between predicted predictability and actual predictability (using various ML methods) is the meta-metric for predicting predictivity.

If the proposed data smashing distance is in general a better metric than Shannon entropy, the absolute value of the coefficient of correlation between validation accuracy and data smashing dis-

tance should be higher than the absolute value of the coefficient of correlation between validation accuracy and other entropy measures. The reason to use absolute value is because data smashing and entropy are a reversal order indicator related to dataset predictivity. The higher the DS distance is, the higher the predictivity. Meanwhile, the lower the entropy is, the higher the predictivity.

We explore this meta metric by doing the following. Our experiment and implementation are conducted in two cases: simulated data and daily industrial sector price time series data (of GICS, Global Industry Classification Standard, level III sectors data). We aim to validate both the simulated dataset and real-world dataset, and whether data smashing distance is in general a better metric than Shannon entropy.

5.5.1 Toy Problem

For the simulated data, we designed and used a Probabilistic Finite State Automata (PFSA) to generate stochastic time series. The design is the following: the PFSA has two states, and transitions probabilities are shown in the Figure.5.3. The two states have the probability distribution of (0/70% and 1/30%) and (0/30% and 1/70%) respectively. If these two states are merged, the probability will average to around 50%/50%. We used this PFSA to generate the stochastic data stream with the length of 10k symbols for our toy problem (we later refer to it as the toy problem stream).

We generated multiple toy problem data streams. Each time a toy problem stream was created, we calculated its Shannon entropy and its data smashing distance to a flat white noise. In addition, each time one a toy problem stream was generated, we applied machine learning algorithms to train and calculate validation accuracy. The toy problem data stream was divided into 90% for training and 10% for validation. In this experiment, we used a stochastic prediction algorithm as the baseline [111], and the Receiver Operator Curve (ROC) area as the metric to calculate its validation accuracy.

We repeated this process 100 times, and ended up with three vectors with the length of 100. vector entries are 100 trial toy problem streams Shanon Shannon entropy, data smashing distance to

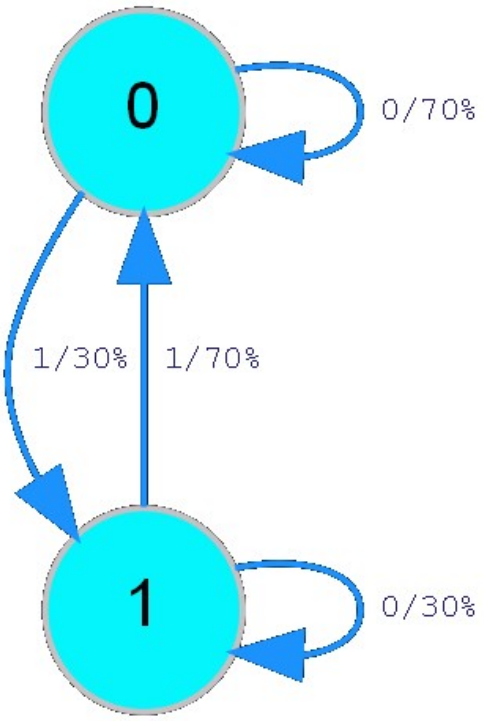


Figure 5.3: Designed PFSA machine that generates the toy problem stream. It has two states, state 0 has probability distribution of 0/70% 1/30%, and state 1 0/30% 1/70%. States transition is shown in the figure.

a flat white noise and validation accuracy. At this point, we calculate the coefficient of correlation between entropy measures validation accuracy and the coefficient of correlation between data smashing distance validation accuracy.

As shown in Table 5.1., the data smashing distance correlates the accuracy(0.1881) better than other entropy.

Table 5.1: Data-smashing metric and other entropy measures' absolute value of coefficient of correlation

Metric	**Absolute value of Coefficient of Correlation**
Data Smashing Distance	0.1881
Shannon entropy	0.0651
Lempel-Ziv entropy	0.0157
Permutation entropy	0.0037
Sample entropy	0.036

This result shows that the data smashing distance is a better metric than other entropy measures in this control experiment. In addition, as mentioned before, the average probability of two PFSA states is 50%/50%. In Figure.5.4., the histogram of 100 toy problem streams Shannon entropy is plotted and the values are clustering around 0.9812, meaning that if Shannon entropy is used to pre-gauge the level of predictivity of a system, it will be viewed as relatively chaotic, as shown in Figure. 5.4. This is clearly not true, given that the toy problem time series are created using a known PFSA model, and they possess a certain temporal pattern.

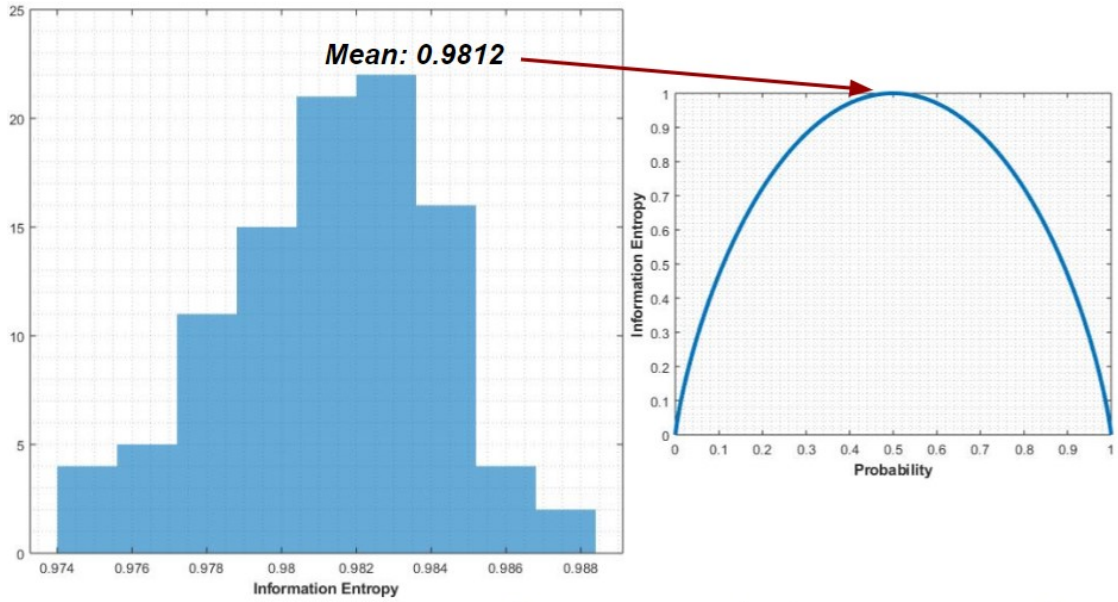


Fig. 4 Histogram of 100 toy problem series, the average is 0.9822. This shows time series is very chaotic.

Figure 5.4: Histogram of 100 toy problem series Shannon entropy value, the average is 0.9822. This shows time series is very chaotic.

Moreover, we plotted the 0/1 ratio of the toy problem streams and their AUC values in Figure 5.5. 0/1 ratio values are around 0.5, which well correlates with the high Shannon entropy value. However, the mean of the AUC value is 0.6, which is definitely not trivial. Thus, in this case, Shannon entropy would not be an ideal metric to use for gauging the predictivity of the dataset before doing any machine learning.

In Figure 5.6, the histogram of the data smashing distance of flat white noise to itself is plotted along with the toy problem streams data smashing distance to flat white noise. Its clearly shown their values are clustering around different value at different magnitude, 0.0025 an 0.0169 respec-

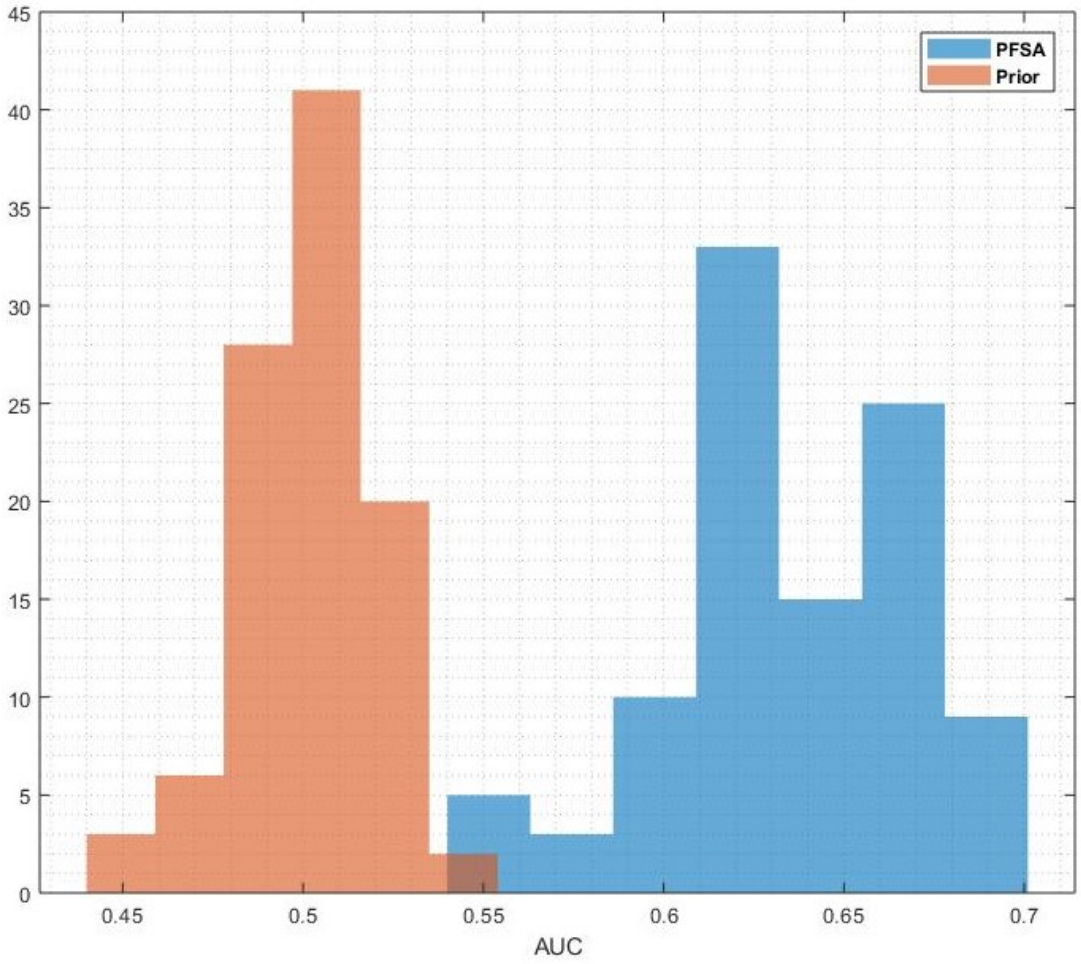


Figure 5.5: AUC value of testset if prior distribution is used(in orange) or PFSA is used(in blue)

tively, this result distinctly revealed the hidden extra information in the dataset detected by our metric compare to using other entropy as metric.

As an alternative to Shannon Entropy, we explored other information metrics. For example, the Lempel-Ziv compression algorithm[113,119,120] uses a dictionary to encode and compress the data. However, if one uses Lempel-Ziv to compress a totally randomly generated data stream, the compression rate will be close to zero due to the randomness despite a lack of valuable information contained. On the other hand, data smashing can not only pick up the temporal pattern but it can also recognize a coin toss if the smash distance is very close to 0 as shown in the toy problems.

One might argue that because the data smashing mathematical model is also based on PFSA, if one uses data smashing distance as a data predictivity indicator and then PFSA to make prediction,

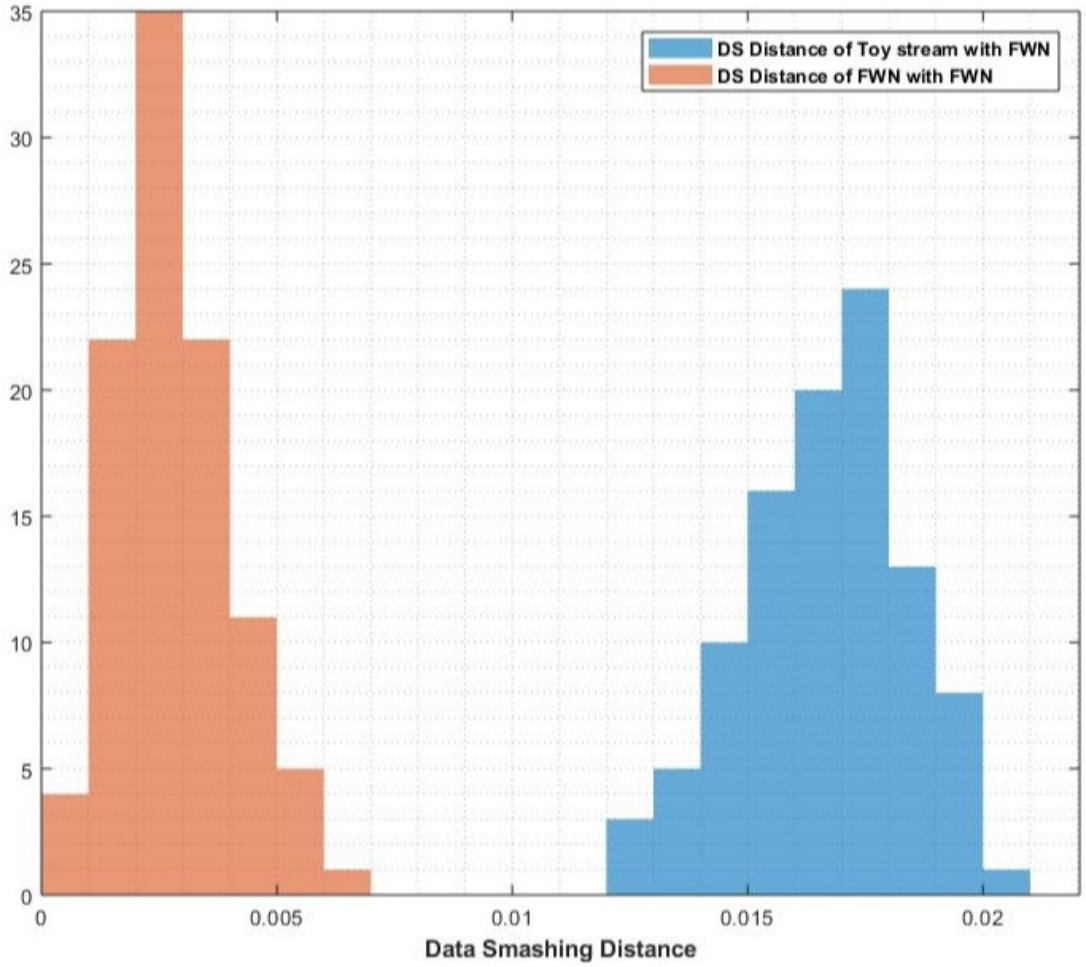


Figure 5.6: Histogram of data smashing distance of flat white noise of itself in orange and data smashing distance of the toy problem data with flat white noise in blue. This shows how data smashing can differentiate pattern data from pure randomness.

it will end up showing a higher correlation in this simulated situation because their methods stem from PFSA.

Does the proposed data smashing metric have an advantage over other entropy metrics when other machine learning methods are used to make predictions? In the following section, we used real world data and other machine learning approaches to give an affirmative answer.

5.5.2 Global Market Data

For industrial sector prices time series, the data starts from January 2, 2002 and ends on July 14, 2017. Each time series has 16,380 observations. GICS is a tiered, hierarchical industry classifi-

cation system, wherein the companies are classified quantitatively and qualitatively in the system. Each different levels represents one tier of the industry level. The higher the tier, the more detailed the data in that sublevel. For example, level one includes Technology, Energy etc. and level three includes more detailed Energy Equipment Services, and Oil, Gas Consumable Fuels. The Bloomberg terminal [114] provides a convenient API to download the industrial sector price time series data.

The time series of price are calculated into returns and are further quantized into binary strings, with the symbol 0, indicating a negative movement in return, and 1, indicating an identical or positive movement in return. We choose level III data, which includes 60 stocks in total as the testing time series. We applied 22 different standard machine learning techniques to train and calculate each models validation accuracy. The training procedure is the same for all 22 methods we applied and the details are as follows.

After quantization, we arbitrarily choose a training window size of five bits. With every five bits as the training set, the machine learning algorithm classified whether the next bit forward in time is 0 (drop) or 1 (raise). After training, we used five-fold cross validation to get model validation accuracy. Meanwhile, we calculated five different entropy measures for each time series: Shannon entropy, Lempel-Ziv entropy, the proposed data smashing distance, permutation entropy and sample entropy. From here, like we did in the toy problem, for each machine learning algorithm, we calculated the correlation between time series validation accuracy and their corresponding entropy calculated from the five entropy measures.

In Table 5.2, the 22 machine learning methods are listed in the first and third columns, and their correlations are printed on the right. We highlighted the highest correlation among five entropy measures in red and plotted the histogram of the metrics with the highest correlation out of 22 machine learning methods. From Figure. 5.7, we can clearly see the dominance advantage of our metric compared to the others. Out of the 22 machine learning methods we applied, 14 highest correlations are produced from our metric.

In addition, We added one more dimension to further analyze and compare our new metric.

Table 5.2: Absolute value of 5 metrics coefficients of Correlation for different Machine Learning methods

Machine learning Algo	SE	LZ	DS	Per	Sample
Fine Tree	0.2184	0.2447	***0.2577***	0.1318	0.2316
Fine KNN	0.0508	***0.1559***	0.0571	-0.0356	0.0829
Medium Tree	0.6897	0.2764	***0.7089***	0.1148	0.6534
Medium KNN	***0.2503***	0.1253	0.2329	0.1191	0.2367
Coarse Tree	0.6905	0.2727	***0.7045***	0.0355	0.6901
Coarse KNN	0.5596	0.2607	***0.5738***	0.11	0.2319
Linear Discriminant	0.7798	0.2585	***0.8177***	0.0364	0.7774
Cosine KNN	***0.2448***	0.1274	0.2315	0.1119	0.2319
Quadratic Discriminant	0.5279	0.2908	***0.5771***	0.1340	0.5431
cubic KNN	***0.2513***	0.1218	0.2386	0.112	0.2338
Linear SVM	0.7765	0.2536	***0.8165***	0.0392	0.776
Weighted KNN	***0.2481***	0.1378	0.2361	0.1107	0.2385
Quadratic SVM	0.6754	0.2669	***0.6862***	0.0345	0.6861
Boosted Trees	0.2386	0.2607	***0.2780***	0.1212	0.2481
Cubic SVM	0.3404	0.2026	0.3694	0.076	***0.392***
Bagged Trees	0.3498	0.2813	***0.3903***	0.1298	0.3618
Fine Gaussian SVM	0.2650	0.2620	***0.3003***	0.1482	0.2648
Subspace Discriminant	0.8603	0.2837	***0.9103***	0.1444	0.7978
Medium Gaussian SVM	0.3657	0.2431	***0.4044***	0.1400	0.3749
Subspace KNN	0.0732	0.0655	***0.0816***	-0.0314	0.1036
Coarse Gaussian SVM	0.8003	0.3015	***0.8720***	0.1134	0.7956
RUSBoosted Trees	-0.2442	***0.1990***	-0.1650	-0.0204	0.1372

For all 60 time series, we applied a moving window with a length of 2000 to calculate time series SE and the DS distance within window, and move one step forward at a time. In Figure 5.8. we take the 1st time series as one example, and find that a strong negative correlation(-0.941) between SE and DS can be observed. If SE is close to one, or the DS distance is small towards zero (close to flat white noise), this implies a very chaotic system. Figure. 5.8 also shows that both methods capture the evolution of a systems chaoticness over time.

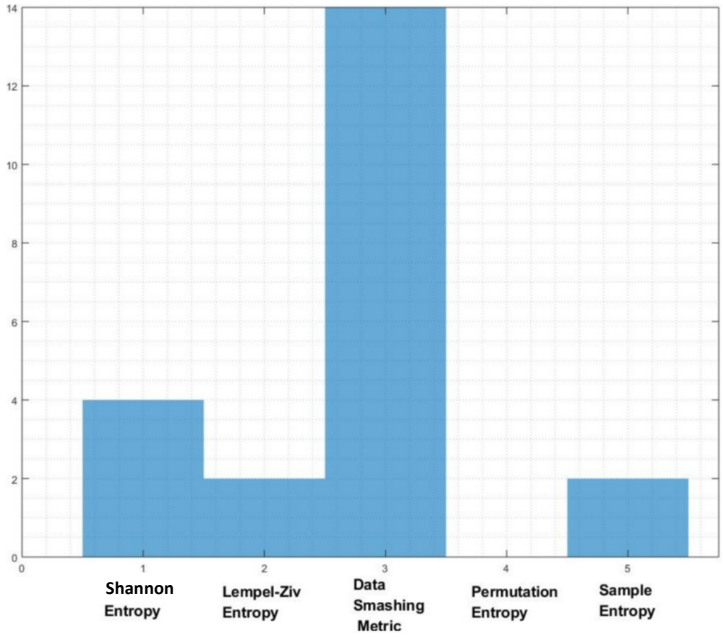


Figure 5.7: Histogram of highest correlation for each metric. Out of 22 machine learning methods, the data smashing metric has 14 highest, Shannon entropy 4, Lempel-Ziv and sample entropy both has 2.

However, if we plot the moving windows SE and DS correlation with accuracy in a histogram in Figure. 5.9, it shows that, on average, the DS distance has a higher correlation with accuracy compared to SE. Even thought the DS distance and IE are highly negatively correlated, the DS distance as a new metric, must capture more of the nuance of the dataset, given its ability to detect temporal memory. Therefore, this proves to be a better metric than SE which is proven by the higher correlation with accuracy.

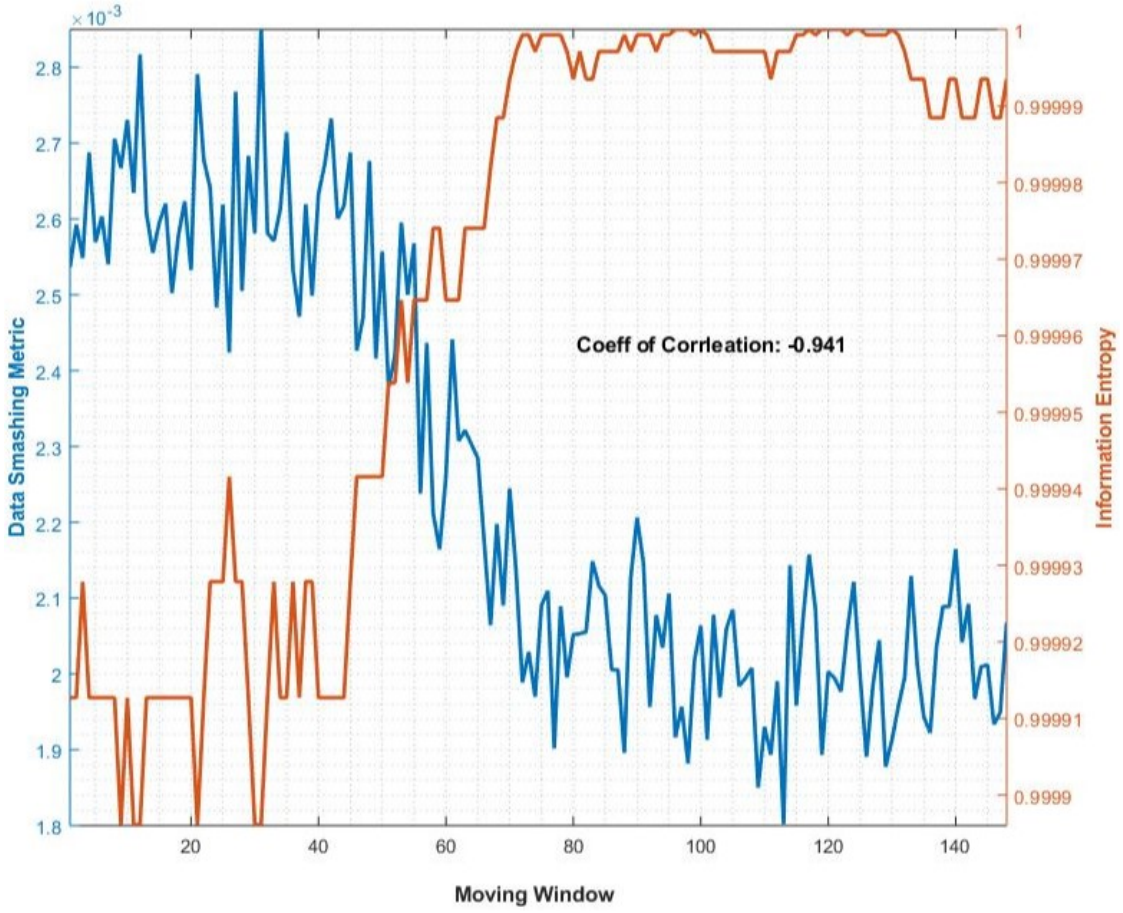


Figure 5.8: Calculated Shannon entropy and data smashing metric within moving windows for 1st time series, it is self-evident that they are highly negatively correlated. However, Data Smashing metric trajectory is more volatile than Shannon entropy. Its picking up more nuance than Shannon entropy.

5.6 Discussion

For our proposed predictivity index, which is based on data smashing, there are no parameters that need to be specified, except for a quantization scheme. The existence of PFSA generators is implicitly assumed, even though this follows the assumption that the time series of interest satisfy the properties of ergodicity, stationarity and have a finite number of states in PFSA. In [13] we argue that any quantized ergodic stationary stochastic process is representable as a probabilistic automata (see [13] Section S-D in the electronic supplementary material).

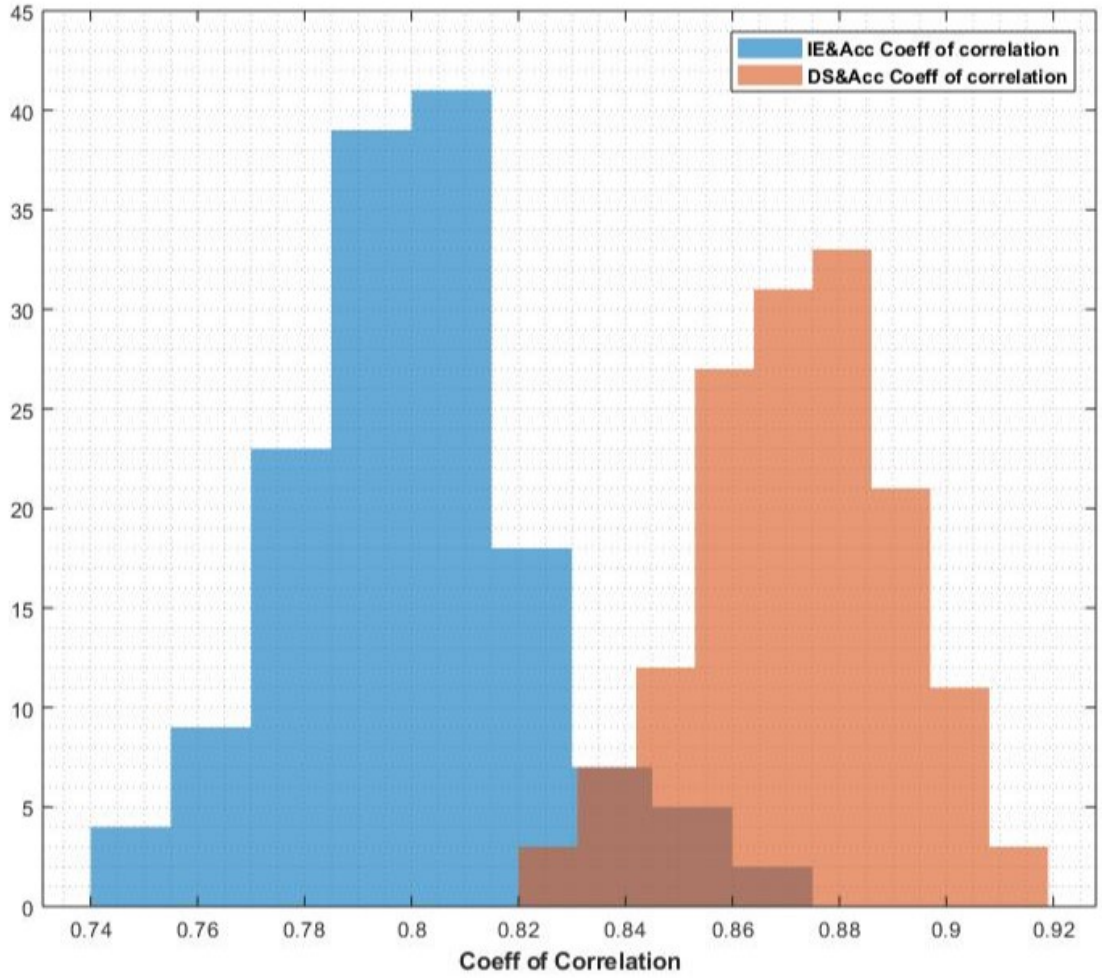


Figure 5.9: Histogram of coefficient of correlation of validation accuracy Shannon entropy and coefficient of correlation of validation accuracy Data Smashing distance of moving windows for 1st time series

Chapter 6: Conclusion

6.1 Conclusions for Chapters

For this dissertation, I successfully implemented our causality inference algorithm to

1. Develop a higher order causality network and successfully used it to capture the second order causal relationship of Bitcoin arbitrage activity between different exchanges

2. Confirm the seismic activities' causal relationship from the middle American Trench to California from a data perspective

3. Achieved state-of-the-art human brain signals classification accuracy without having any domain/prior knowledge in advance. We discovered mice neuron activity patterns using mice brain signals without making any assumptions

4. Detect the causal pattern in the data stream. We successfully invented an index to represent the time series predictivity and showed that our method is the state-of-the-art compared to other entropy measures.

Chapter 1 extended the previous causality network research. A non-parametric model was designed to quantitatively measure and test the level of existence of higher order causality between ergodic stationary weakly dependent symbol sources. Beyond binary test, rejecting or accepting a null hypothesis, our method produces a structure generative model. This allows us to investigate deep the dynamic causal relationship between data flows further. Higher order moment property can also be explored in this fashion, specifically the second order moment, like the cases studies that are illustrated in the experiments section. One can repeat the procedure of inferring the second order to infer higher moment like skewness and kurtosis. In the experiments and applications, we

studied a toy problem to prove the validity of our approach, and applied this method to industrial sector data and calculated the second order causality network to explore the sectors volatility causal relationship. In addition, we applied thid method to study the arbitrage trading activity of Bitcoin between two exchanges. This work allows us to tap deeper into the complex causal relationships of higher order system so that we can better understand the hidden causal mechanism.

Using earthquake catalogs across California and the Middle American Trench, we identified a delayed hidden causality pattern between 250 to 350 weeks in **chapter 2**. This delay had the ROC AUC value of approximately 0.62, with the confidence level above 99%. This result confirms the hypothesis put forward by Raleigh et al (2). We further suggest that a similar analysis can be performed globally to find additional regions bound by causality effects. Ultimately, the complete collection of effects can be considered as the global seismic network.

Since our approach is purely data-driven, we can only nd statistical causality, and not geophysical causation. We cannot, for example, determine the mechanisms that underlie the long-range interactions between the seismogenic structures, and what roles are being played by the elastic seismogenic zone and the viscous asthenosphere that lies beneath. While in one sense such a physics-free approach facilitates our analysis with no knowledge of the material properties involved, and requires no detailed empirical modeling of stress relaxation dynamics, it does become a black-box solution, which gives answers without providing a rst-principles model.

Finally, our analysis shows that accurate earthquake prediction is a non-local prediction problem since prediction is influenced by seismic activity in remote seismic regions and significant improvement in prediction performance is to be expected when these influences are included.

In **Chapter 3**, we present a novel approach to classifying brain signals that outperforms the state of the art techniques as follows:

1. It provides high accuracy independently of the expertise of the user or nature of the data. No preprocessing, feature extraction, or supervision of the algorithm are needed.

   This opens up the BCI science and industry to people from all ages and backgrounds. The algorithm is plug-and-play.

2. It distinguishes similar classes with high accuracy even with modest amounts of unlabeled data.

   This will enable inexpensive consumer-product systems to classify multiple neural states; avoiding the need for specialized hardware.

3. It Is insensitive to user bias or prevailing notions of brain function.

   This will enable BCI to improve despite limitations in our current knowledge. By providing successful observations, it will advance our understanding of brain functions.

4. It works with multiple types of neural signals.

   This will extend the utility and reach of BCI.

5. It is fast to setting up and run, and it can be parallelized.

   This makes it a perfect candidate for real-time applications.

Future work can be explored in two sections. First it is important to further explore why certain *single* signals can capture the features we discovered. If successful, this will shed the light on our deeper understanding of the brain.

Secondly, it is important to explore the data smashing ability to classify time series in more other cases, such as financial time series, etc.. Since it does not require any expert/prior knowledge, data smashing can be applied to any nature of time series in a plug-and-play fashion. This provides researchers the leverage to circumvent the several bottlenecks of modern machine learning.

In **Chapter 4**, we introduced a predictivity index as a new way to measure the level of complexity of the time series. This is a parameter and distribution-free way of calculating dataset predictivity that does not require any expert tuned heuristics. The advantage of this method is demonstrated both in the synthetic and financial dataset by comparing the correlation between validation accuracy and our metric, which is the correlation between other entropy measures.

Even though the proposed predictivity index metric does not always have the highest correlation with performance of a future machine learning method, statistically speaking, its superiority

is evident in both case studies. Its other strength also focuses on circumventing any required expert knowledge and rendering our method as an easy plug-and-play metric. Therefore, this can be readily implemented by anyone in any user case.

6.2 Contribution of Others to This Thesis

I want to thank Professor Rafael Yuste and his Ph.D. student, Shuting Han. They provided the mice brain dataset and offered biological insights into the mice section of the data smashing project in Chapter 3.

I also want to express my gratitude to Professor Francisco Valero-Cuevas and his Ph.D. student, Ali Marjaninejad. From them, I have gained access to the BCI signal of human. Professor Valero-Cuevas and Ali also helped to pre-process the dataset and discussed the meaning of the machine learning algorithm results. In the end, Ali also helped with the writing in Chapter 3.

I want to thank Professor Yehuda Ben-Zion and Anghel Marian for their help with Chapter 2. They supplemented the knowledge in the earthquake domain, refining the papers concept, and proofreading. Even though Professor Yehuda let half-way through, they are open minded geologists who are open-minded to the idea of global earthquakes causality system. For this, Im forever grateful.

I want to acknowledge Professor Agostino Capponi for his help with regards to the work around the higher-order causality network. Professor Capponi helped with the concept validation, as well as with writing and proofreading.

I also want to thank my colleagues Rob Kwiatkowski, Oscar Chang, Chad DeChant, and Philippe Martin Wyder for proofreading my paper.

Bibliography

1. https://en.wikipedia.org/wiki/Correlation_does_not_imply_causation

2. https://en.wikipedia.org/wiki/Granger_causality

3. Granger, Clive WJ, Bwo-Nung Huangb, and Chin-Wei Yang. "A bivariate causality between stock prices and exchange rates: evidence from recent Asianflu." The Quarterly Review of Economics and Finance 40.3 (2000): 337-354.

4. Hiemstra, Craig, and Jonathan D. Jones. "Testing for linear and nonlinear Granger causality in the stock pricevolume relation." The Journal of Finance 49.5 (1994): 1639-1664.

5. Brovelli, Andrea, et al. "Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by Granger causality." Proceedings of the National Academy of Sciences 101.26 (2004): 9849-9854.

6. Kamiski, Maciej, et al. "Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance." Biological cybernetics85.2 (2001): 145-157.

7. Roebroeck, Alard, Elia Formisano, and Rainer Goebel. "Mapping directed influence over the brain using Granger causality and fMRI." Neuroimage 25.1 (2005): 230-242.

8. Woniak, Tomasz. "Granger-causal analysis of VARMA-GARCH models." (2012).

9. Pantelidis, Theologos, and Nikitas Pittis. "Testing for Granger causality in variance in the presence of causality in mean." Economics Letters 85.2 (2004): 201-207.

10. Cheung, Yin-Wong, and Lilian K. Ng. "A causality-in-variance test and its application to financial market prices." Journal of Econometrics 72.1-2 (1996): 33-48.

11. Ishanu Chattopadhyay, Causality Networks,arXiv:1406.6651 [cs.LG]

12. Hafner, Christian M., and Helmut Herwartz. "Testing for causality in variance using multivariate GARCH models." Annales d'Economie et de Statistique (2008): 215-241.

13. Chattopadhyay, Ishanu, and Hod Lipson. "Data smashing: uncovering lurking order in data." Journal of The Royal Society Interface 11.101 (2014): 20140826.

14. E. G. Baek and W. A. Brock, A general test for nonlinear granger causality: Bivariate model, Jan. 1992.

15. C. Hiemstra and C. Kramer, Nonlinearity and endogeneity in macro-asset pricing, in International Monetary Fund Working Paper, 1995.

16. I. Asimakopoulos, D. Ayling, and W. M. Mahmood, Non-linear granger causality in the currency futures returns, Economics Letters, vol. 68, no. 1, pp. 25 30, 2000.

17. C. Hiemstra and J. D. Jones, Testing for Linear and Nonlinear Granger Causality in the Stock Price-Volume Relation, The Journal of Finance, vol. 49, no. 5, pp. 16391664, 1994.

18. Woniak, Tomasz. "Granger-causal analysis of GARCH models: a Bayesian approach." Econometric Reviews 37.4 (2018): 325-346.

19. Comte, Fabienne, and Offer Lieberman. "SecondOrder Noncausality in Multivariate GARCH Processes." Journal of Time Series Analysis 21.5 (2000): 535-557.

20. Zhang, Kun, and Aapo Hyvarinen. "Source separation and higher-order causal analysis of MEG and EEG." arXiv preprint arXiv:1203.3533 (2012).

21. Chattopadhyay, Ishanu, and Hod Lipson."Abductive learning of quantized stochastic processes with probabilistic finite automata."Phil. Trans.R.Soc.A 371.

22. Fine, Shai, Yoram Singer, and Naftali Tishby. "The hierarchical hidden Markov model: Analysis and applications." Machine learning 32.1 (1998): 41-62

23. https://www.bloomberg.com/professional/solution/bloomberg-terminal/

24. https://en.wikipedia.org/wiki/Entropy_(information_theory)

25. https://www.businessinsider.com/this-is-what-hedge-funds-are-buying-right-now-2015-8

26. https://en.wikipedia.org/wiki/Global_Industry_Classification_Standard

27. Nakamoto, Satoshi. "Bitcoin: A peer-to-peer electronic cash system." (2008).

28. Madan, Isaac, Shaurya Saluja, and Aojia Zhao. "Automated bitcoin trading via machine learning algorithms."

29. Makarov, Igor and Schoar, Antoinette, Trading and Arbitrage in Cryptocurrency Markets (April 30, 2018). Available at SSRN: https://ssrn.com/abstract=3171204

30. Marshall, Ben R. and Nguyen, Nhut Hoang and Visaltanachoti, Nuttawat, Bitcoin Liquidity (June 12, 2018). Available at SSRN: https://ssrn.com/abstract=3194869

31. Chan, Kalok. "A further analysis of the leadlag relationship between the cash market and stock index futures market." The Review of Financial Studies 5.1 (1992): 123-152.

32. Karpoff, Jonathan M. "The relation between price changes and trading volume: A survey." Journal of Financial and quantitative Analysis 22.1 (1987): 109-126.

33. Denis, François. "PAC learning from positive statistical queries." International Conference on Algorithmic Learning Theory. Springer, Berlin, Heidelberg, 1998.

34. Auer, Peter, Robert C. Holte, and Wolfgang Maass. "Theory and applications of agnostic PAC-learning with small decision trees." Machine Learning Proceedings 1995. 1995. 21-29.

35. Long, Philip M., and Lei Tan. "PAC learning axis-aligned rectangles with respect to product distributions from multiple-instance examples." Machine Learning 30.1 (1998): 7-21.

36. Raleigh, C. B., Sieh, K., Sykes, L. R., Anderson, D. L. (1982), Forecasting southern California earthquakes. Science, 217(4565), 1097-1104.

37. Sadovsky, M. A., Nersesov, I. L. (1974), Forecasts of earthquakes on the basis of complex geophysical features. In: T. Rikitake (Editor), Focal Processes and the Prediction of Earthquakes. Tectonophysics, 23(3), 247-255.

38. Granger, C. W. J. (1980), Testing for causality: A personal viewpoint. Journal of Economic Dynamics and Control, 2(4), 329-352.

39. Astiz, L., Kanamori, H., Eissler, H. (1987), Source Charactristics of earthquakes in the Michoacan seismic gap in Mexico. Bulletin of the seismological society of America, 77(4): 1326-1346.

40. Toppozada, T., Branum, D. (2004), California earthquake history. Annals of Geophysics, 47(2-3), Special Issue, SI, 509-522.

41. Schorlemmer, D., Wiemer, S. (2005), Microseismicity data forecast rupture area. Nature, 434(7037), 1086.

42. Nishenko, S. P., Bollinger, G. A. (1990), Forecasting damaging earthquakes in the Central and Eastern United States. Science, 249 (4975), 1412-1416.

43. Rouet-Leduc, B., Hulbert, C., Lubbers, N., Barros, K., Humphreys, C. J., Johnson, P. A. (2017), Machine learning predicts laboratory earthquakes. Geophysical Research Letters, 44(18), 9276-9282.

44. Sykes, L. R., Jaume, S. C. (1990), Seismic activity on neighboring faults as a long-term precursor to large earthquakes in the San Francisco Bay area. Nature, 348(6302), 595-599.

45. Ward, S. N., Goes, S. D. B. (1993), How regularly do earthquakes recur? A synthetic seismicity model for the San Andreas Fault. Geophysical Research Letters, 20(19), 2131-2134.

46. Knopoff, L., Levshina, T., Keilis-Borok, V. I., Mattoni, C. (1996), Increased long-range intermediate-magnitude earthquake activity prior to strong earthquakes in California. Journal of Geophysical Research-Solid Earth, 101(B3), 5779-5796.

47. Bowman, D. D., Ouillon, G., Sammis, C. G., Sornette, A., Sornette, D. (1998), An observational test of the critical earthquake concept. Journal of Geophysical Research-Solid Earth, 103(B10), 24359-24372.

48. Prejean, S. G., Hill, D. P., Brodsky, E. E., Hough, S. E., Johnston, M. J. S., Malone, S. D., Oppenheimer, D. H., Pitt, A. M., Richards-Dinger, K. B. (2004), Remotely triggered seismicity on the United States west coast following the Mw 7.9 Denali Fault Earthquake. Bulletin of the Seismological Society of America, 94(6B), S348-S359.

49. Robinson, R. (2004), Potential earthquake triggering in a complex fault network: the northern South Island, New Zealand. Geophysical Journal Intentional, 159(2), 734-748.

50. Gerstenberger, M. C., Wiemer, S., Jones, L. M., Reasenberg, P. A. (2005), Real-time forecasts of tomorrows earthquakes in California. Nature, 435(7040), 328-331.

51. Toda, S., Lin, J., Meghraoui, M., Stein, R. S. (2008), 12 May 2008 M = 7.9 Wenchuan, China, earthquake calculated to increase failure stress and seismicity rate on three major fault systems. Geophysical Research Letters, 35(17), L17305.

52. Marzocchi, W., Lombardi, A. M. (2009), Real-time forecasting following a damaging earthquake. Geophysical Research Letters, 36(21), L21302.

53. Barbot, S., Lapusta, N., Avouac, J. P. (2012), Under the Hood of the Earthquake Machine: Toward Predictive Modeling of the Seismic Cycle. Science, 336(6082), 707-710.

54. Toda, S., Stein, R. S. (2013), The 2011 M=9.0 Tohoku oki earthquake more than doubled the probability of large shocks beneath Tokyo. Geophysical Research Letters, 40(11), 2562-2566.

55. Huang, J. P., Wang, X. A., Zhao, Y., Xin, C., Xiang, H. (2018), Large earthquake magnitude prediction in Taiwan based on deep learning neural network. Neural Network World, 28(2), 149-160.

56. Kariche, J., Meghraoui, M., Timoulali, Y., Cetin, E., Toussaint, R. (2018), The Al Hoceima earthquake sequence of 1994, 2004 and 2016: Stress transfer and poroelasticity in the Rif and Alboran Sea region. Geophysical Journal International, 212(1), 42-53.

57. Wang, Q., Guo Y., Yu, L., Li, P. Earthquake prediction based on spatio-temporal data mining: an LSTM network approach. IEEE Transactions on Emerging Topics in Computing, to appear.

58. United states geological survey earthquake database. http://earthquake.usgs.gov/earthquakes/ Accessed, 2017-02-01.

59. Chattopadhyay, I. (2014), Causality networks. http://arxiv.org/abs/1406.6651.

60. Baek, E. G., Brock, A. W. (1992), A general test for nonlinear granger causality: Bivariate model. Technical Report, Korean Development Institute and University of Wisconsin-Madison.

61. Hiemstra, C., Jones, J. D. (1994), Testing for Linear and Nonlinear Granger Causality in the Stock Price-Volume Relation. The Journal of Finance, 49(5), 1639-1664.

62. Valiant, L. G. (1984), A theory of the learnable. Communications of the ACM, 27(11), 1134-1142.

63. Gutenberg, B., Richter, C. F. (1954), Seismicity of the earth and associated phenomena. Princeton University Press, Princeton, NJ, US.

64. Chattopadhyay, I., Lipson, H. (2013), Abductive learning of quantized stochastic processes with probabilistic finite automata. Philosophical Transactions of the Royal Society A-Mathematical, Physical and Engineering Sciences, 371(1984), Article Number, UNSP 20110543

65. Shaw, B. E. (2004). Variation of large elastodynamic earthquakes on complex fault systems. Geophysical Research Letters, 31(18), Article Number, L18609.

66. Anghel, M., Ben-Zion, Y., Rico-Martinez, R. (2004). Dynamical system analysis and forecasting of deformation produced by an earthquake fault. Pure and Applied Geophysics, 161(9-10), 2023-2051.

67. DeVries, P., F. Viegas, M. Wattenberg, and B. Meade (2018), Deep learning of aftershock patterns following large earthquakes, Nature 560, 7720, 632..

68. Ross, Z. E., Yue, Y., Meier, M.-A., Hauksson, E., and T. H. Heaton (2018). PhaseLink: A Deep Learning Approach to Seismic Phase Association, J. Geophys. Res.-Solid Earth, in review, [arXiv:1809.02880].

69. Ross, Z. E., Meier, M.-A., Hauksson, E., and T. H. Heaton (2018). Generalized Seismic Phase Detection with Deep Learning, Bull. Seismol. Soc. Am., doi: 10.1785/0120180080 [arXiv:1805.01075].

70. S. Sanei and J. A. Chambers, EEG signal processing. John Wiley Sons, 2013.

71. N. Liang and L. Bougrain, Decoding finger flexion from band-specific ECoG signals in humans, Front. Neurosci, 2012.

72. C. Mehring et al., Comparing information about arm movement direction in single channels of local and epicortical field potentials from monkey and human motor cortex, J. Physiol., vol. 98, no. 4, pp. 498506, 2004.

73. T. Aflalo et al., Decoding motor imagery from the posterior parietal cortex of a tetraplegic human, Science (80-.)., vol. 348, no. 6237, pp. 906910, 2015.

74. C. Klaes et al., Hand Shape Representations in the Human Posterior Parietal Cortex, J. Neurosci., vol. 35, no. 46, pp. 1546615476, 2015.

75. G. Schalk et al., Decoding two-dimensional movement trajectories using electrocorticographic signals in humans, J. Neural Eng., vol. 4, no. 3, p. 264, 2007.

76. I. Chattopadhyay and H. Lipson, Data smashing: Uncovering lurking order in data, J. R. Soc. Interface, vol. 11, no. 101, p. 20140826, 2014.

77. J. Sanders and E. Kandrot, CUDA by Example: An Introduction to General-Purpose GPU Programming, Portable Documents. Addison-Wesley Professional, 2010.

78. Xu, Fei, and Joshua B. Tenenbaum. "Word learning as Bayesian inference." Psychological review 114.2 (2007): 245.

79. Perfors, Amy, and Joshua Tenenbaum. "Learning to learn categories." Cognitive Science Society, 2009.

80. Smith, Linda B., et al. "Object name learning provides on-the-job training for attention." Psychological Science 13.1 (2002): 13-19.

81. Kemp, Charles, Amy Perfors, and Joshua B. Tenenbaum. "Learning overhypotheses with hierarchical Bayesian models." Developmental science 10.3 (2007): 307-321.

82. Fei-Fei, Li, Rob Fergus, and Pietro Perona. "One-shot learning of object categories." IEEE transactions on pattern analysis and machine intelligence 28.4 (2006): 594-611.

83. Fe-Fei, Li. "A Bayesian approach to unsupervised one-shot learning of object categories." Proceedings Ninth IEEE International Conference on Computer Vision. IEEE, 2003.

84. Socher, Richard, et al. "Zero-shot learning through cross-modal transfer." Advances in neural information processing systems. 2013.

85. Altae-Tran, Han, et al. "Low data drug discovery with one-shot learning." ACS central science 3.4 (2017): 283-293.

86. Rahimi, Abbas, et al. "Hyperdimensional computing for noninvasive brain-computer interfaces: Blind and one-shot classification of EEG error-related potentials." 10th EAI Int. Conf. on Bio-inspired Information and Communications Technologies. No. CONF. 2017.

87. Burrello, Alessio, et al. "One-shot learning for iEEG seizure detection using end-to-end binary operations: Local binary patterns with hyperdimensional computing." 2018 IEEE Biomedical Circuits and Systems Conference (BioCAS). IEEE, 2018.

88. Liu, Huan, and Hiroshi Motoda, eds. Feature extraction, construction and selection: A data mining perspective. Vol. 453. Springer Science Business Media, 1998.

89. Guyon, Isabelle, et al., eds. Feature extraction: foundations and applications. Vol. 207. Springer, 2008.

90. Katajamaa, Mikko, and Matej Orei. "Data processing for mass spectrometry-based metabolomics." Journal of chromatography A 1158.1-2 (2007): 318-328.

91. Marjaninejad, Ali, Babak Taherian, and Francisco J. Valero-Cuevas. "Finger movements are mainly represented by a linear transformation of energy in band-specific ECoG signals." 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2017.

92. Jones, Allan R., Caroline C. Overly, and Susan M. Sunkin. "The Allen brain atlas: 5 years and beyond." Nature Reviews Neuroscience 10.11 (2009): 821.

93. Sunkin, Susan M., et al. "Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system." Nucleic acids research 41.D1 (2012): D996-D1008.

94. Aflalo, Tyson, et al. "Decoding motor imagery from the posterior parietal cortex of a tetraplegic human." Science 348.6237 (2015): 906-910.

95. Maasoumi, Esfandiar, and Jeff Racine. "Entropy and predictability of stock market returns." Journal of Econometrics 107.1-2 (2002): 291-312.

96. Pincus, Steve. "Approximate entropy (ApEn) as a complexity measure." Chaos: An Interdisciplinary Journal of Nonlinear Science 5.1 (1995): 110-117.

97. Richman, Joshua S., and J. Randall Moorman. "Physiological time-series analysis using approximate entropy and sample entropy." American Journal of Physiology-Heart and Circulatory Physiology 278.6 (2000): H2039-H2049.

98. Li, Xiaoli, Gaoxian Ouyang, and Douglas A. Richards. "Predictability analysis of absence seizures with permutation entropy." Epilepsy research 77.1 (2007): 70-74.

99. Molgedey, Lutz, and Werner Ebeling. "Local order, entropy and predictability of financial time series." The European Physical Journal B-Condensed Matter and Complex Systems 15.4 (2000): 733-737.

100. Song, Chaoming, et al. "Limits of predictability in human mobility." Science 327.5968 (2010): 1018-1021.

101. Lu, Xin, et al. "Approaching the limit of predictability in human mobility." Scientific reports 3 (2013): 2923.

102. Ding, Guoru, et al. "On the limits of predictability in real-world radio spectrum state dynamics: From entropy theory to 5G spectrum sharing." IEEE Communications Magazine 53.7 (2015): 178-183.

103. Sinatra, Roberta, and Michael Szell. "Entropy and the predictability of online life." Entropy 16.1 (2014): 543-556.

104. Lempel, Abraham, and Jacob Ziv. "On the complexity of finite sequences." IEEE Transactions on information theory 22.1 (1976): 75-81.

105. Vitanyi, Paul MB, and Ming Li. An introduction to Kolmogorov complexity and its applications. Vol. 34. No. 10. Heidelberg: Springer, 1997.

106. Krumme, Coco, et al. "The predictability of consumer visitation patterns." Scientific reports 3 (2013): 1645.

107. Raidl, Ale. "Estimating the fractal dimension, K 2-entropy, and the predictability of the atmosphere." Czechoslovak Journal of Physics 46.4 (1996): 293-328.

108. Manis, George, M. D. Aktaruzzaman, and Roberto Sassi. "Bubble entropy: an entropy almost free of parameters." IEEE Transactions on Biomedical Engineering 64.11 (2017): 2711-2718.

109. Chattopadhyay, Ishanu, and Hod Lipson. "Computing entropy rate of symbol sources a distribution-free limit theorem." Information Sciences and Systems (CISS), 2014 48th Annual Conference on. IEEE, 2014.

110. Chattopadhyay, Ishanu, and Hod Lipson. "Data smashing: uncovering lurking order in data." Journal of The Royal Society Interface 11.101 (2014): 20140826.

111. Chattopadhyay, Ishanu, and Hod Lipson. "Abductive learning of quantized stochastic processes with probabilistic finite automata." Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 371.1984 (2013): 20110543.

112. https://en.wikipedia.org/wiki/Generalization_error

113. Rytter, Wojciech. "Application of LempelZiv factorization to the approximation of grammar-based compression." Theoretical Computer Science 302.1-3 (2003): 211-222.

114. https://www.bloomberg.com/professional/solution/bloomberg-terminal/

115. Lohr, Steve. "The age of big data." New York Times 11.2012 (2012).

116. https://en.wikipedia.org/wiki/Machine_learning

117. Cesa-Bianchi, Nicolo, Alex Conconi, and Claudio Gentile. "On the generalization ability of on-line learning algorithms." IEEE Transactions on Information Theory 50.9 (2004): 2050-2057.

118. Shannon, Claude Elwood. "A mathematical theory of communication." Bell system technical journal 27.3 (1948): 379-423.

119. Farach, Martin, and Mikkel Thorup. "String matching in lempelziv compressed strings." Algorithmica 20.4 (1998): 388-404.

120. Seroussi, Gadiel, and Abraham Lempel. "Lempel-ziv compression scheme with enhanced adapation." U.S. Patent No. 5,243,341. 7 Sep. 1993.

121. https://www.brainyquote.com/quotes/heraclitus_107157

122. http://observatory.brain-map.org/visualcoding

Appendix: Global earthquake activity causality plot

Global earthquakes causality plots for each individual cluster ROC>=0.55

(Youtube link: https://youtu.be/dKemjirvwXc)

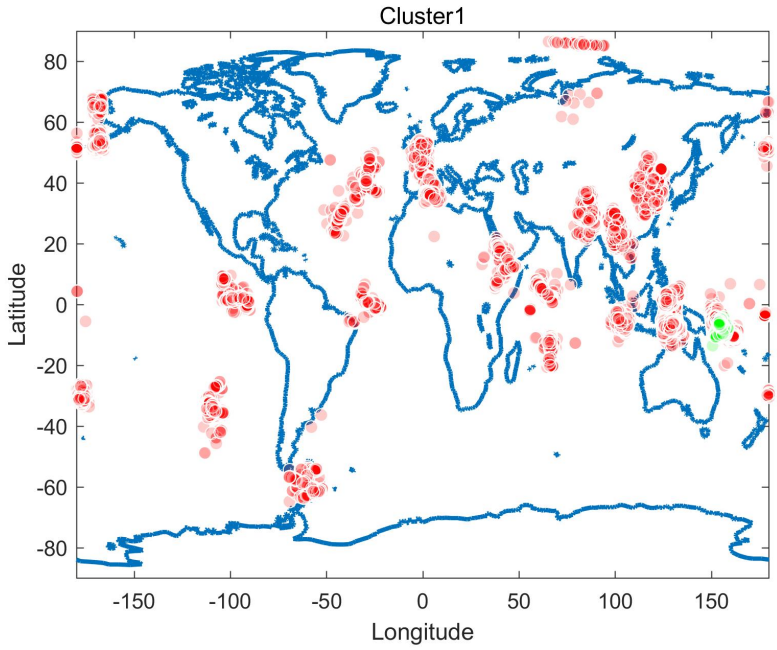


Figure 1: Global earthquakes causal relationship for target cluster1, highlighted in green, all other driving areas highlighted in red

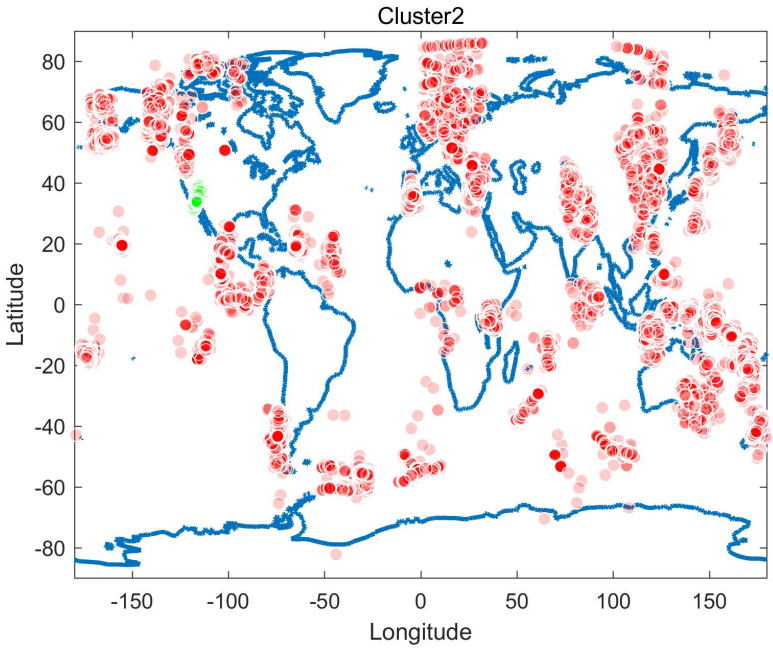


Figure 2: Global earthquakes causal relationship for target cluster2, highlighted in green, all other driving areas highlighted in red

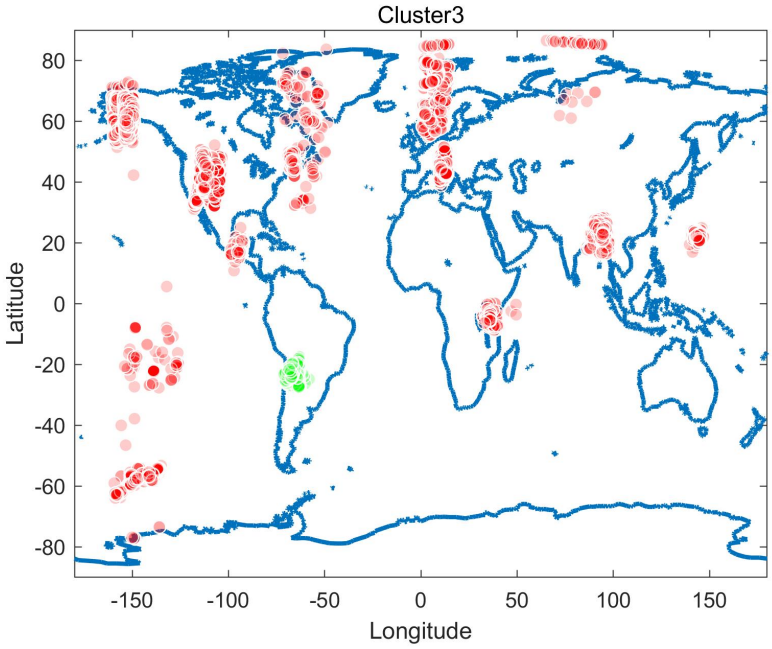


Figure 3: Global earthquakes causal relationship for target cluster3, highlighted in green, all other driving areas highlighted in red

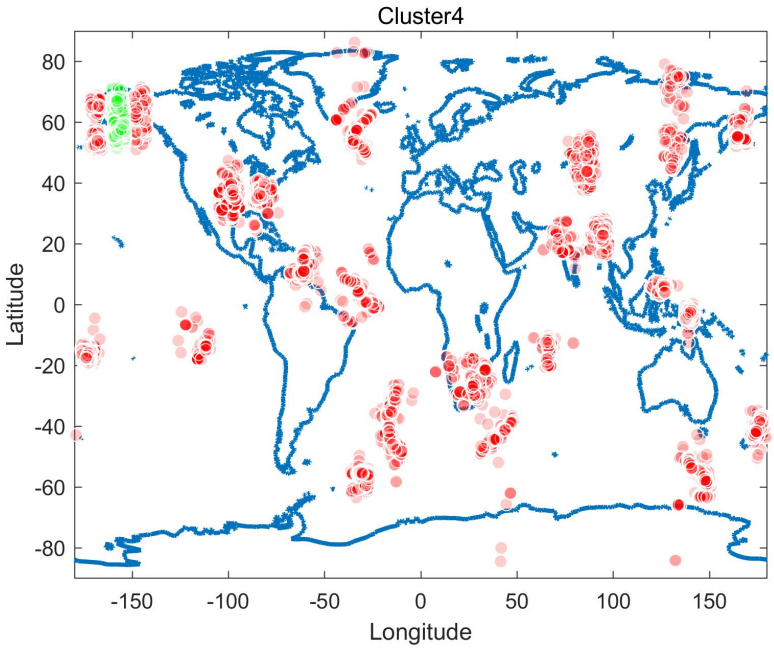


Figure 4: Global earthquakes causal relationship for target cluster4, highlighted in green, all other driving areas highlighted in red

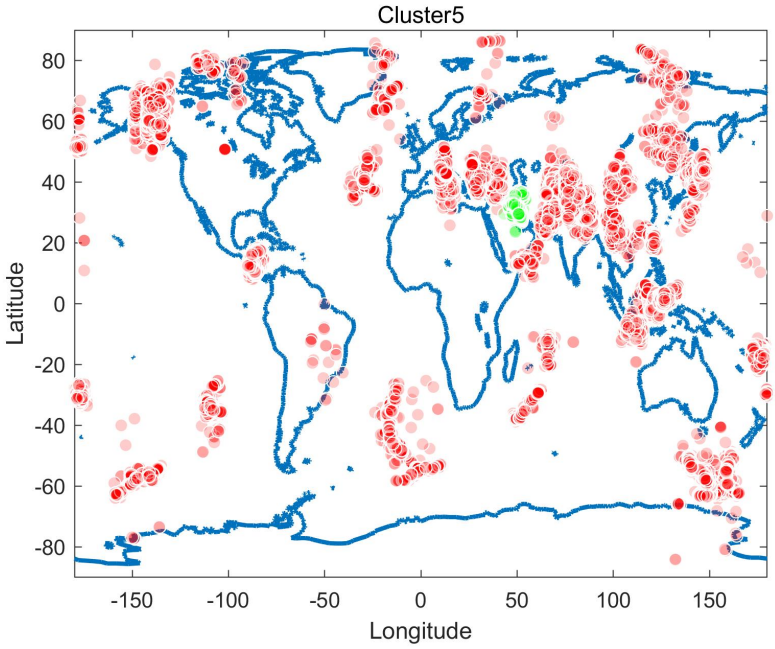


Figure 5: Global earthquakes causal relationship for target cluster5, highlighted in green, all other driving areas highlighted in red

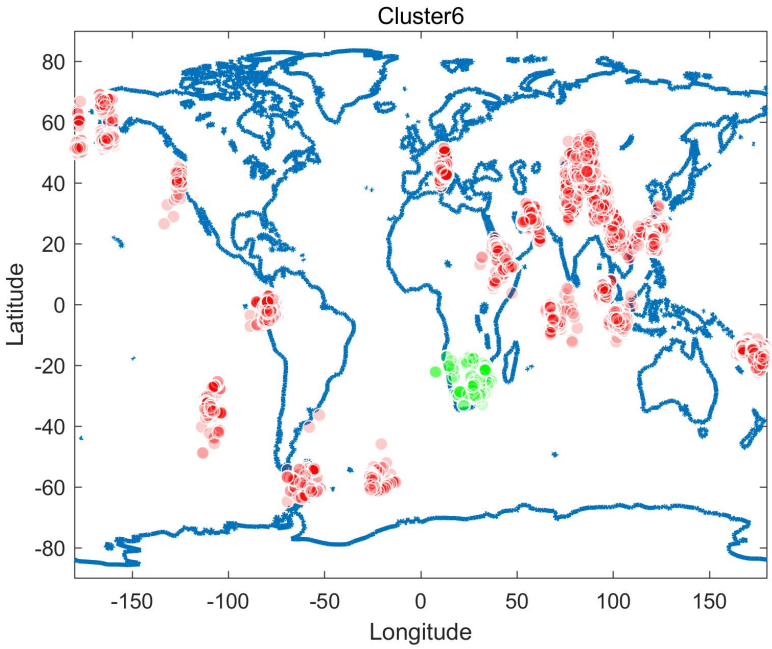


Figure 6: Global earthquakes causal relationship for target cluster6, highlighted in green, all other driving areas highlighted in red

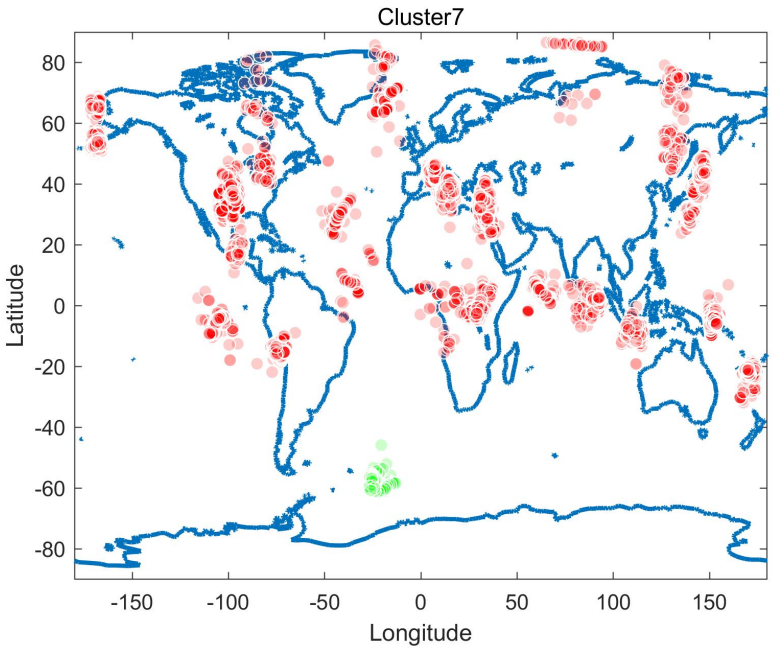


Figure 7: Global earthquakes causal relationship for target cluster7, highlighted in green, all other driving areas highlighted in red

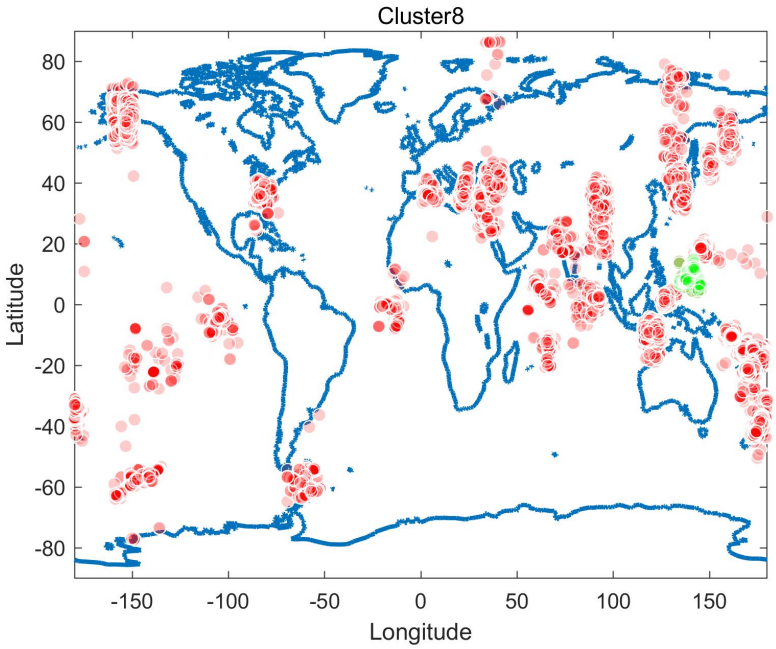


Figure 8: Global earthquakes causal relationship for target cluster8, highlighted in green, all other driving areas highlighted in red

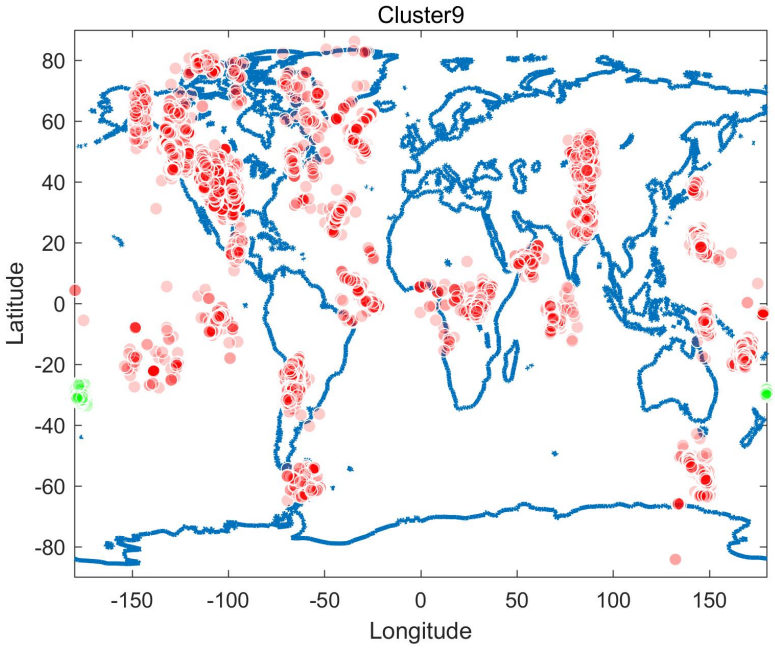


Figure 9: Global earthquakes causal relationship for target cluster9, highlighted in green, all other driving areas highlighted in red

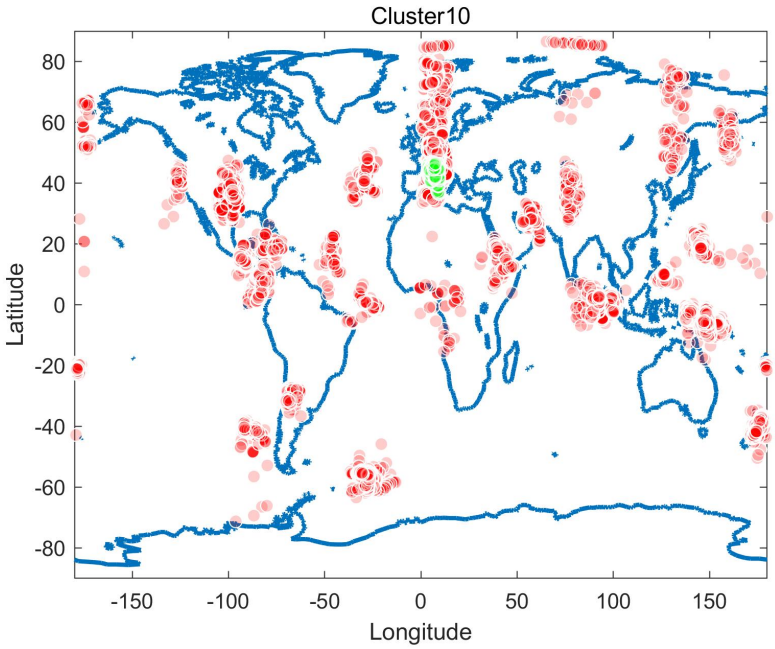


Figure 10: Global earthquakes causal relationship for target cluster10, highlighted in green, all other driving areas highlighted in red

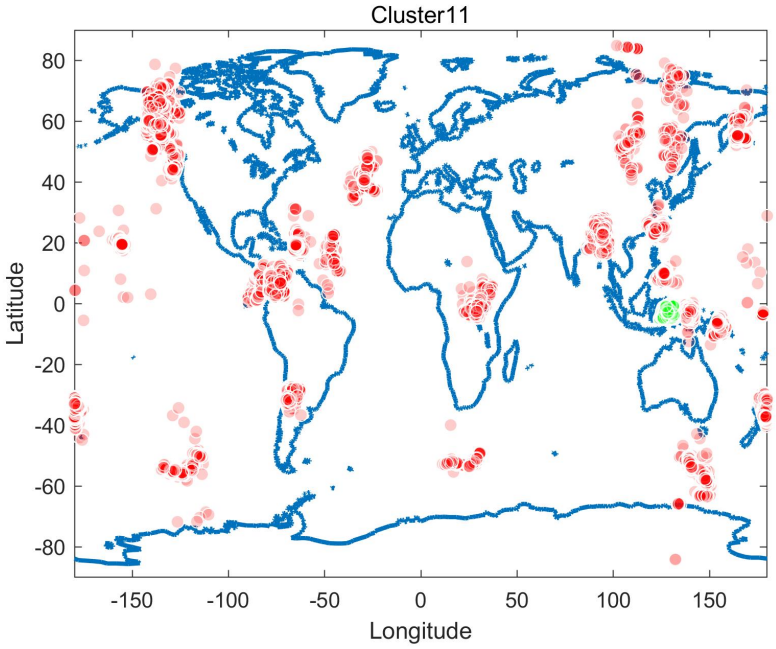


Figure 11: Global earthquakes causal relationship for target cluster11, highlighted in green, all other driving areas highlighted in red

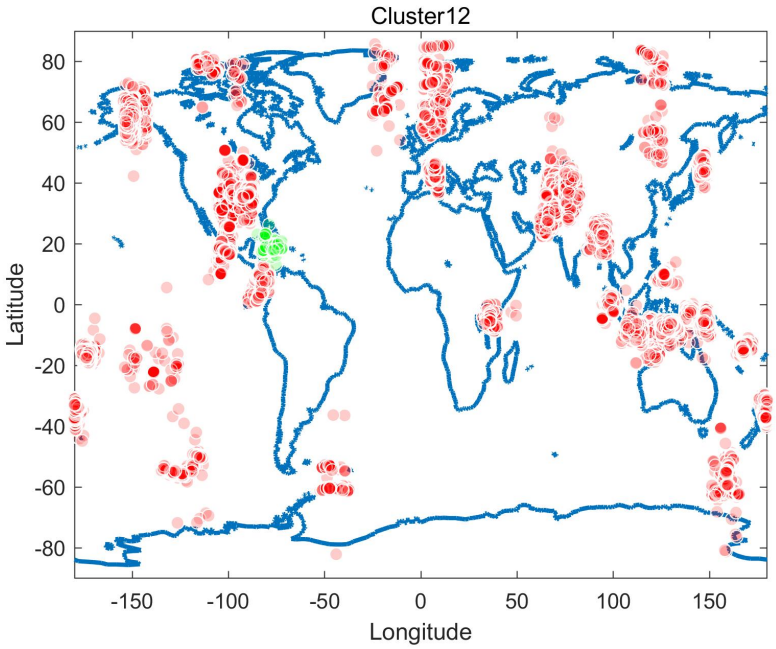


Figure 12: Global earthquakes causal relationship for target cluster12, highlighted in green, all other driving areas highlighted in red

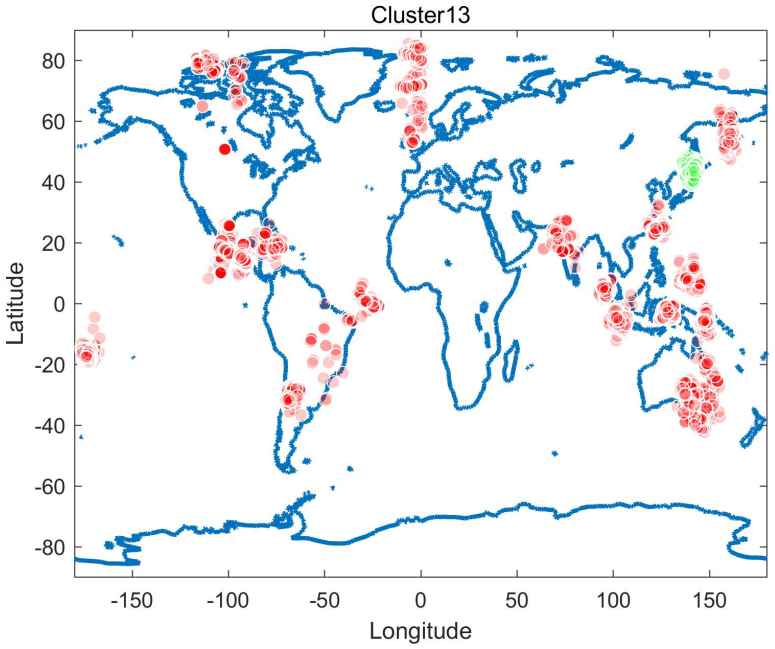


Figure 13: Global earthquakes causal relationship for target cluster13, highlighted in green, all other driving areas highlighted in red

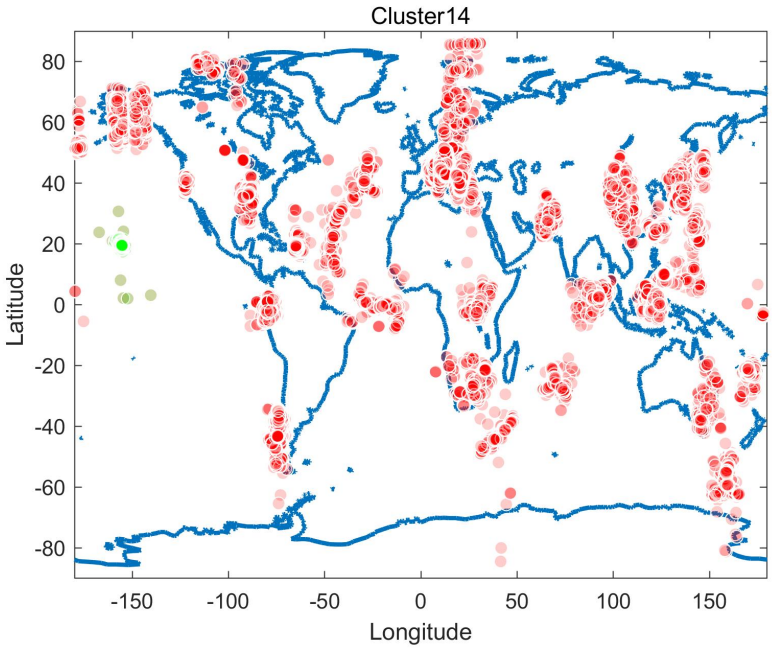


Figure 14: Global earthquakes causal relationship for target cluster14, highlighted in green, all other driving areas highlighted in red

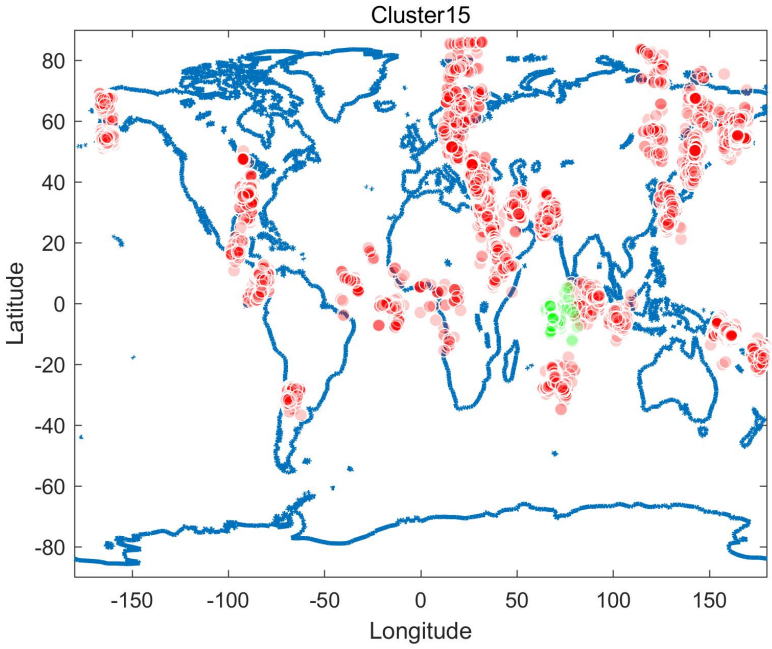


Figure 15: Global earthquakes causal relationship for target cluster15, highlighted in green, all other driving areas highlighted in red

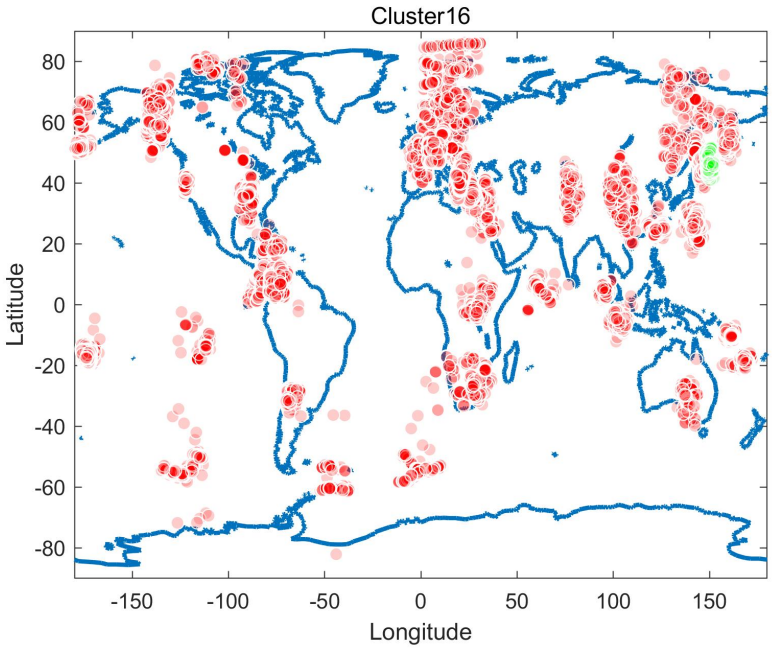


Figure 16: Global earthquakes causal relationship for target cluster16, highlighted in green, all other driving areas highlighted in red

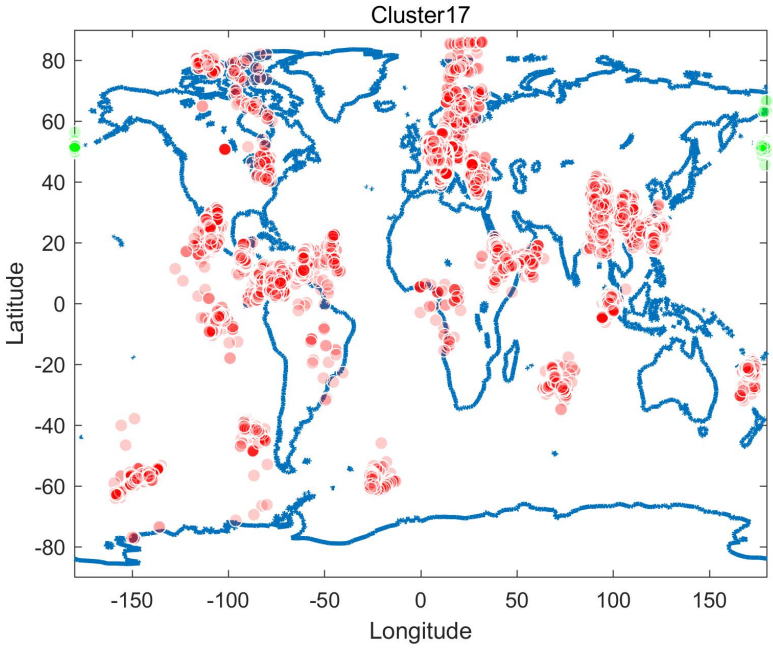


Figure 17: Global earthquakes causal relationship for target cluster17, highlighted in green, all other driving areas highlighted in red

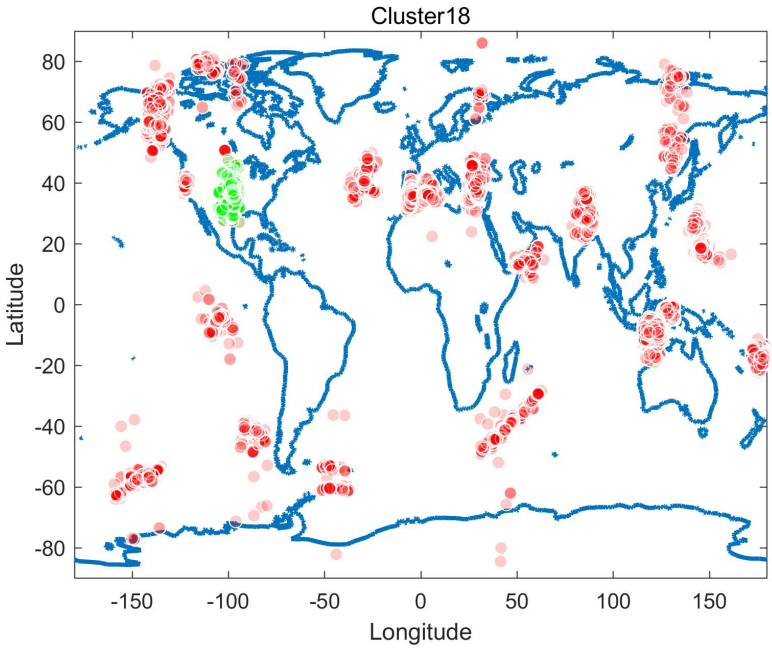


Figure 18: Global earthquakes causal relationship for target cluster18, highlighted in green, all other driving areas highlighted in red

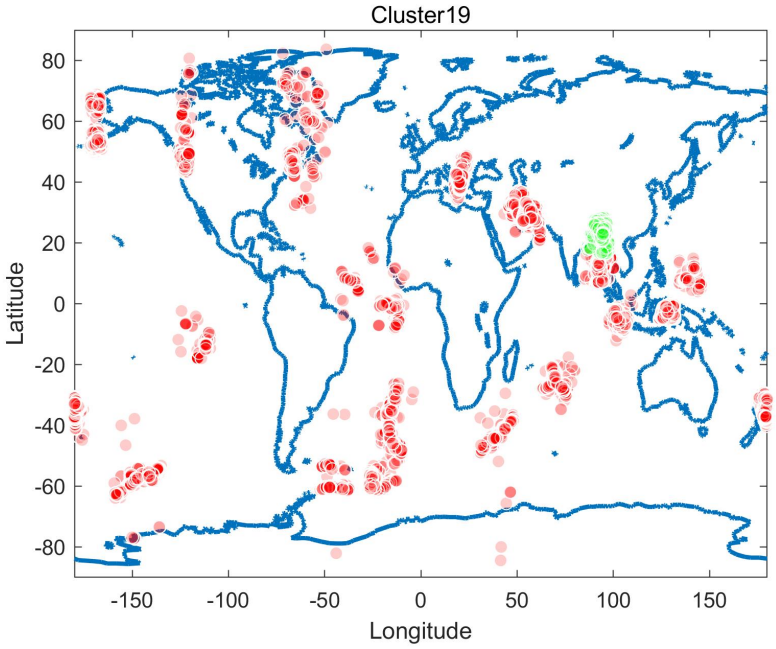


Figure 19: Global earthquakes causal relationship for target cluster19, highlighted in green, all other driving areas highlighted in red

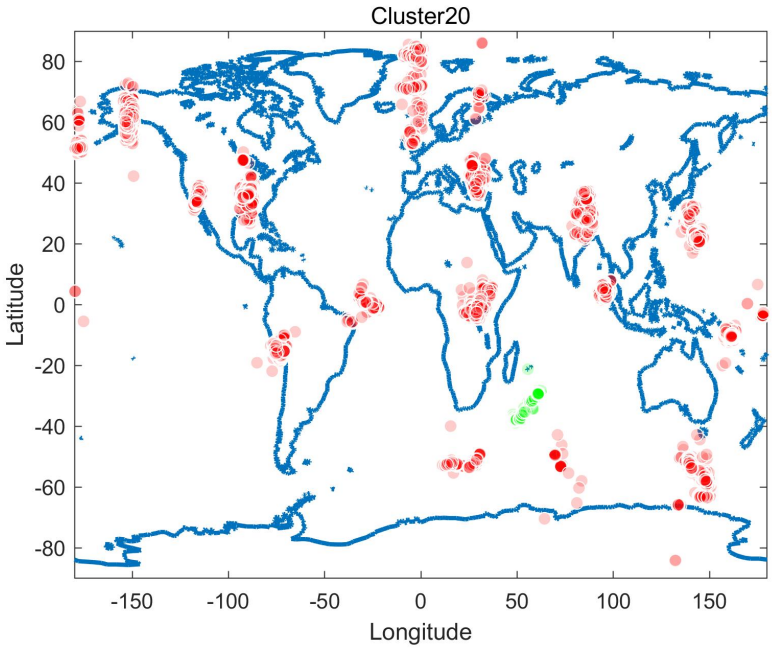


Figure 20: Global earthquakes causal relationship for target cluster20, highlighted in green, all other driving areas highlighted in red

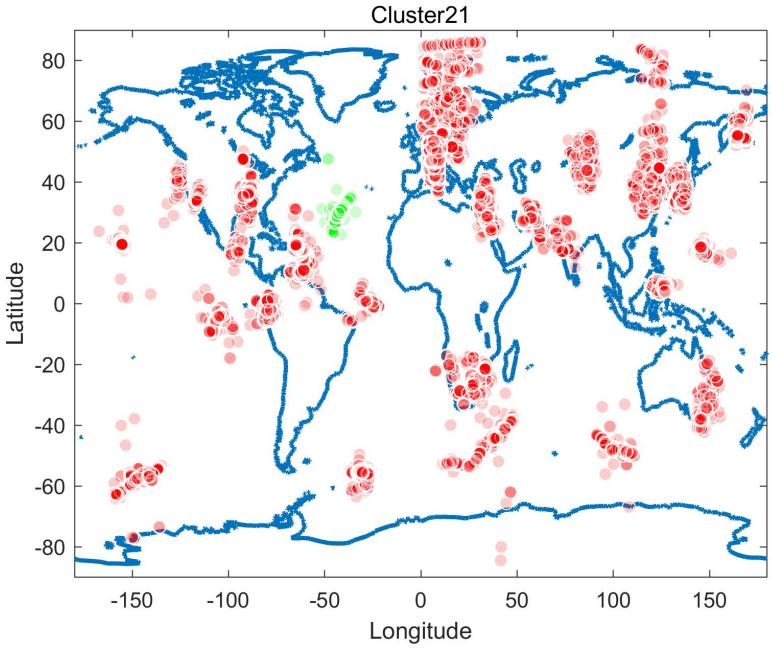


Figure 21: Global earthquakes causal relationship for target cluster21, highlighted in green, all other driving areas highlighted in red

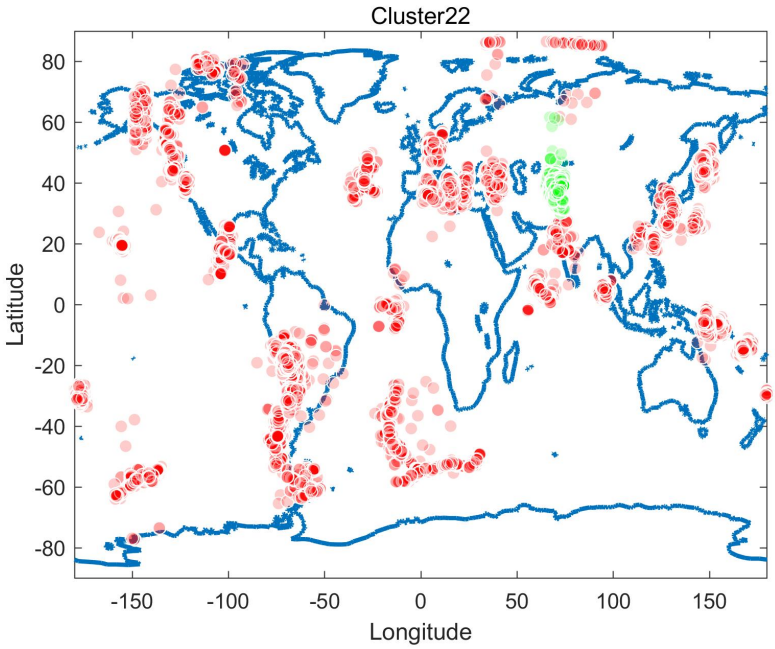


Figure 22: Global earthquakes causal relationship for target cluster22, highlighted in green, all other driving areas highlighted in red

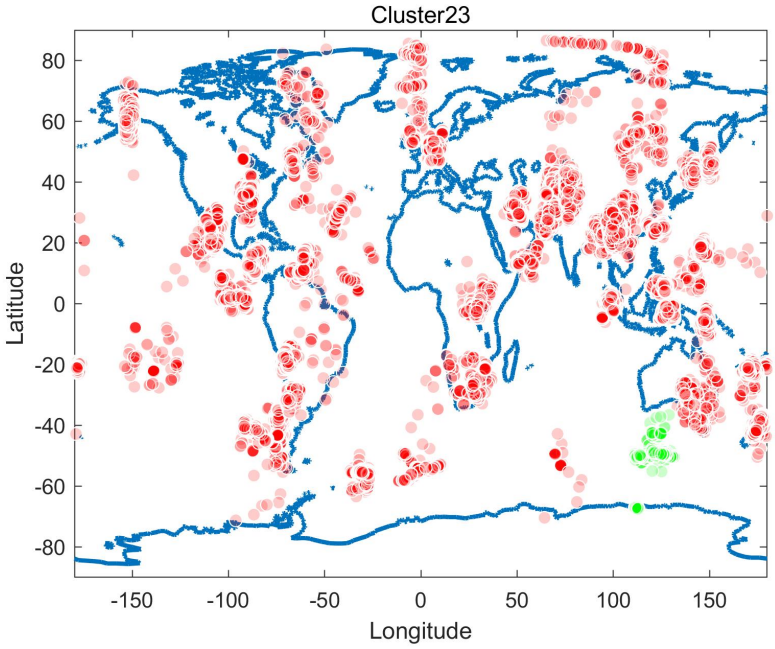


Figure 23: Global earthquakes causal relationship for target cluster23, highlighted in green, all other driving areas highlighted in red

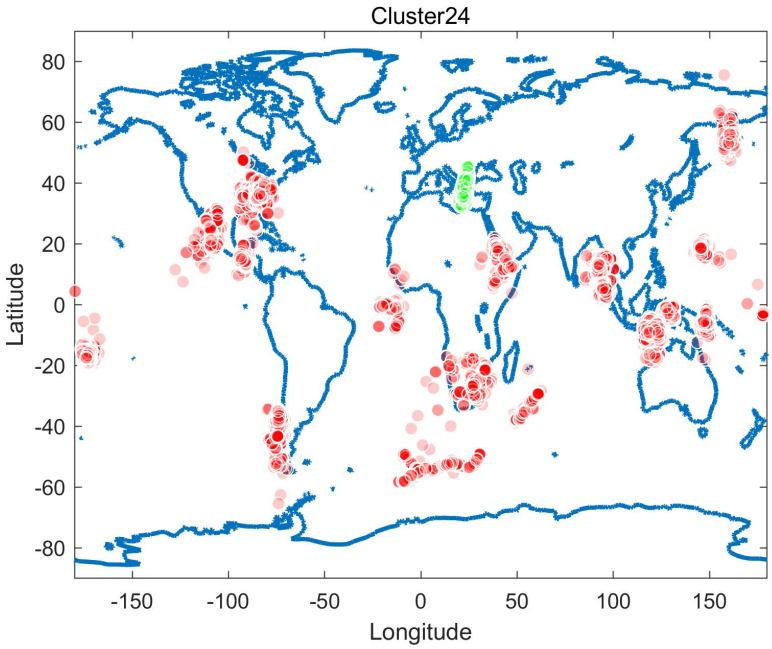


Figure 24: Global earthquakes causal relationship for target cluster24, highlighted in green, all other driving areas highlighted in red

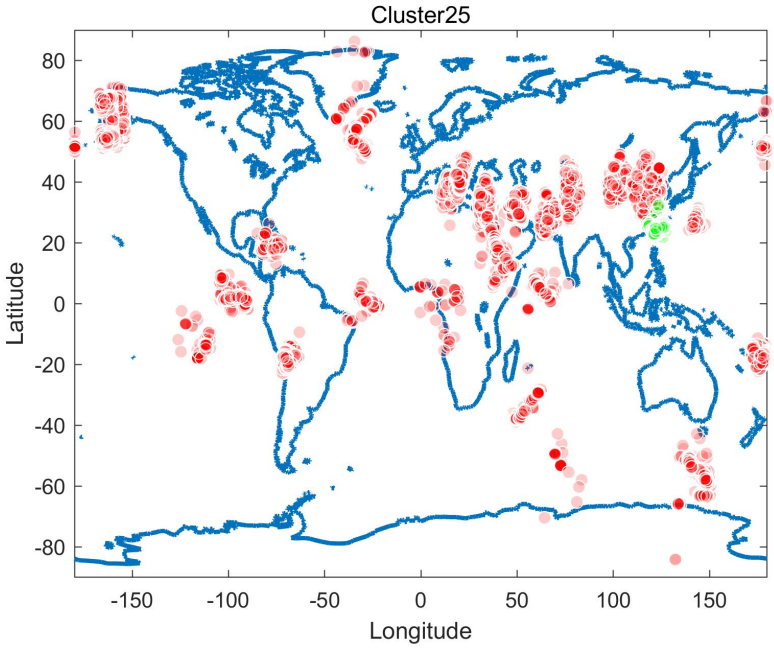


Figure 25: Global earthquakes causal relationship for target cluster25, highlighted in green, all other driving areas highlighted in red

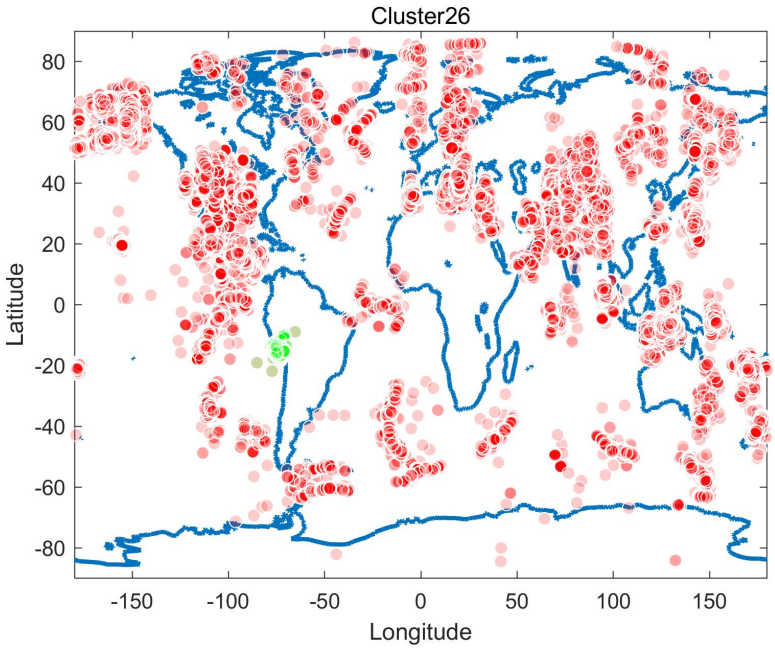


Figure 26: Global earthquakes causal relationship for target cluster26, highlighted in green, all other driving areas highlighted in red

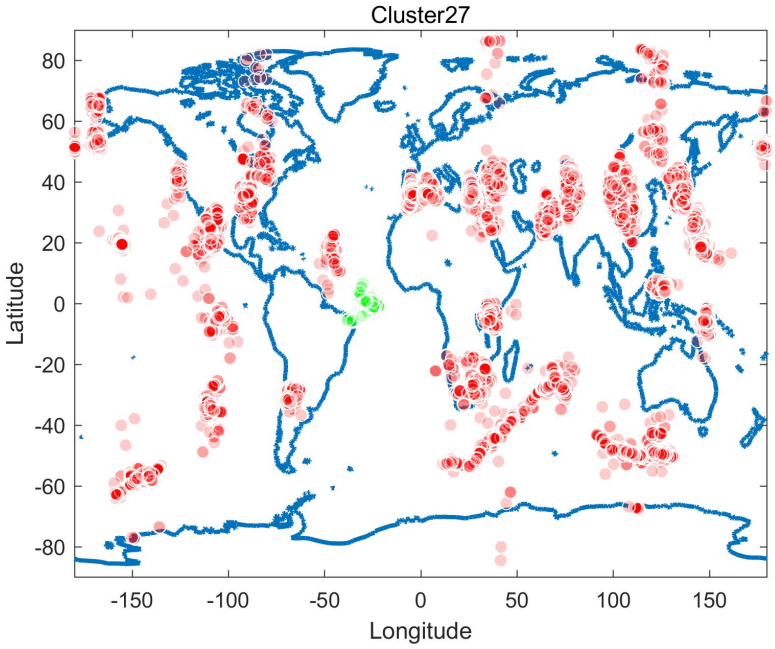


Figure 27: Global earthquakes causal relationship for target cluster27, highlighted in green, all other driving areas highlighted in red

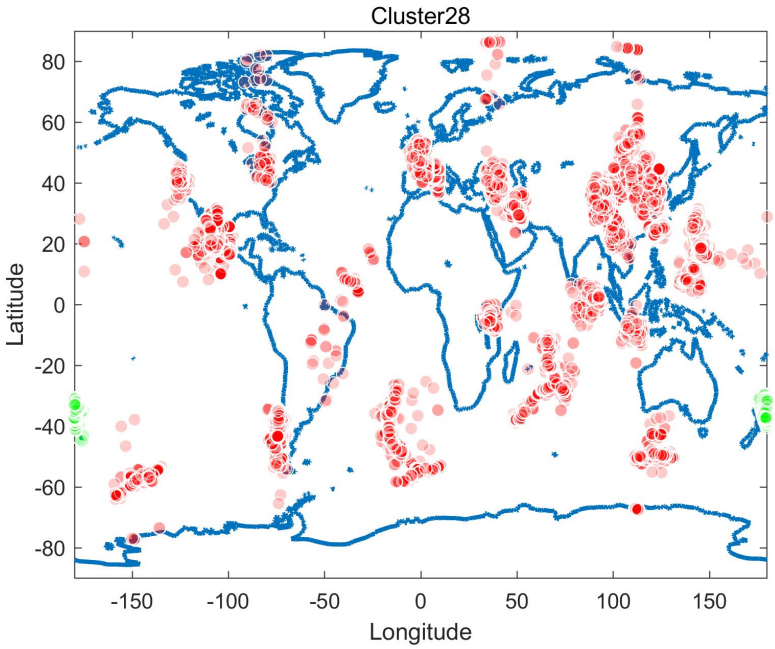


Figure 28: Global earthquakes causal relationship for target cluster28, highlighted in green, all other driving areas highlighted in red

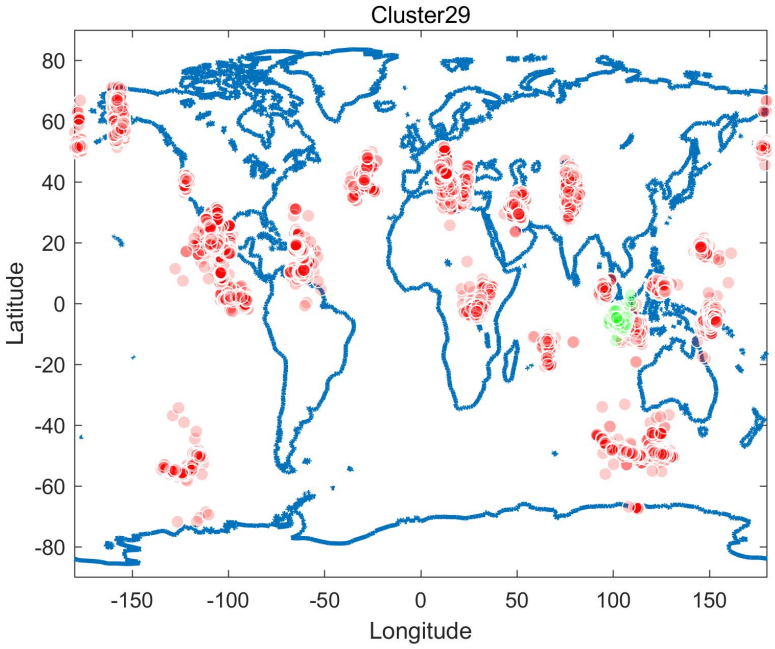


Figure 29: Global earthquakes causal relationship for target cluster29, highlighted in green, all other driving areas highlighted in red

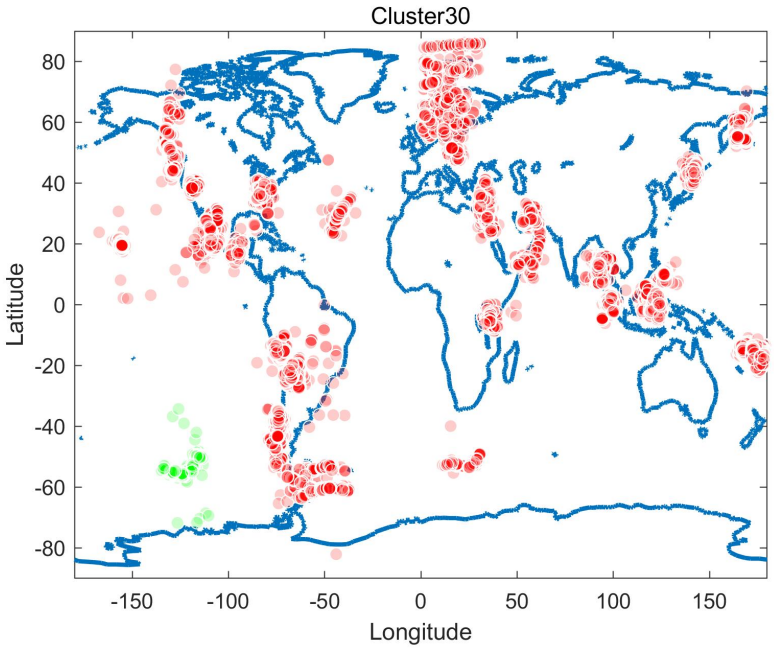


Figure 30: Global earthquakes causal relationship for target cluster30, highlighted in green, all other driving areas highlighted in red

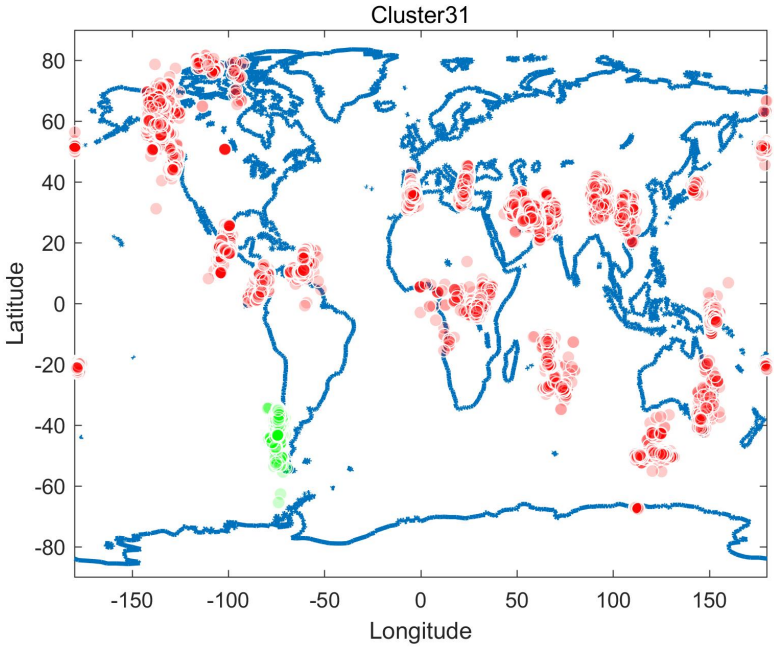


Figure 31: Global earthquakes causal relationship for target cluster31, highlighted in green, all other driving areas highlighted in red

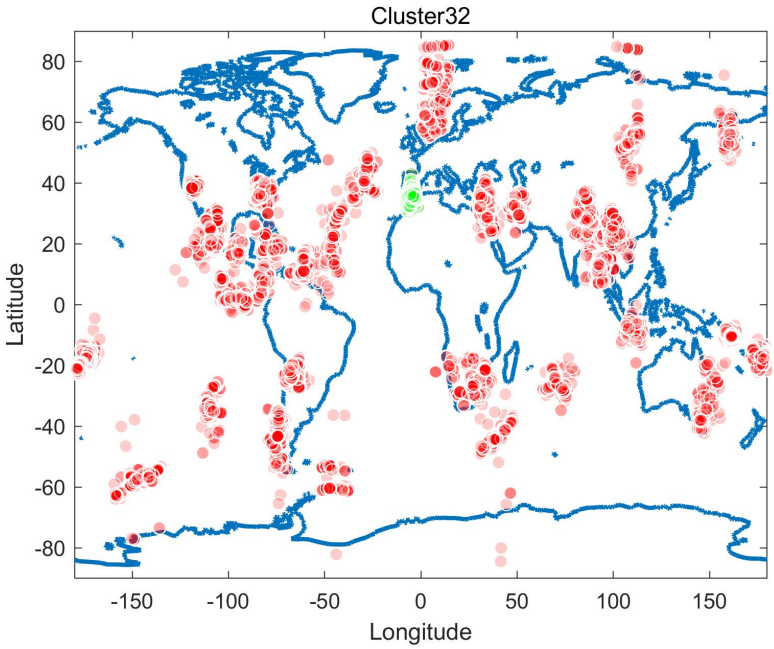


Figure 32: Global earthquakes causal relationship for target cluster32, highlighted in green, all other driving areas highlighted in red

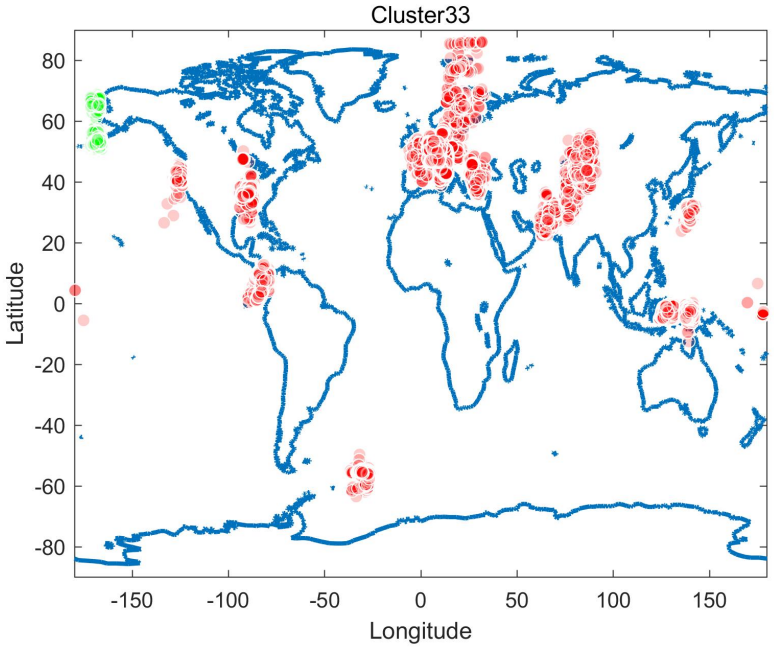


Figure 33: Global earthquakes causal relationship for target cluster33, highlighted in green, all other driving areas highlighted in red

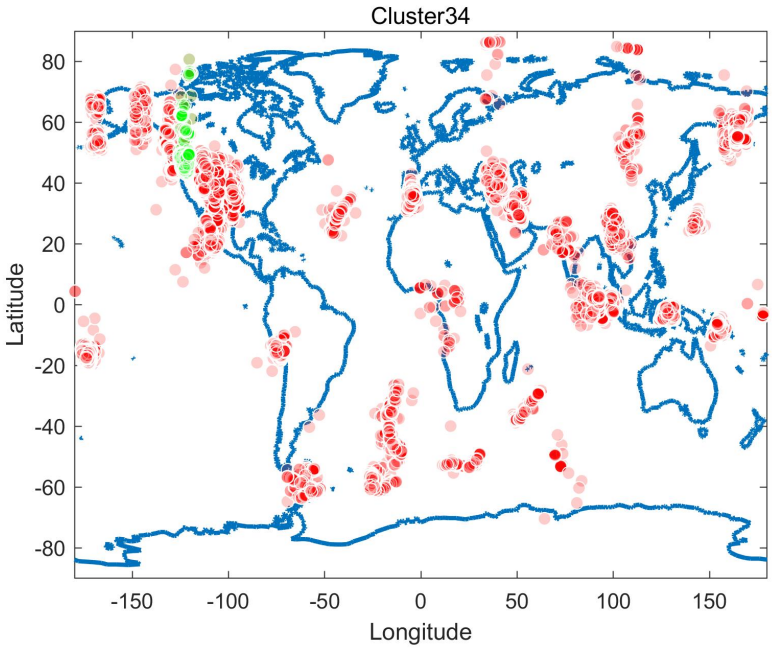


Figure 34: Global earthquakes causal relationship for target cluster34, highlighted in green, all other driving areas highlighted in red

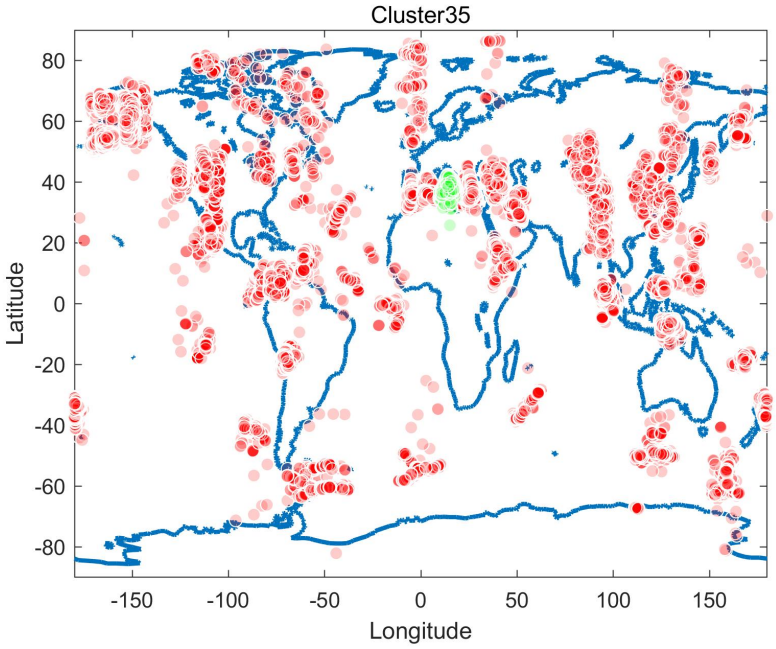


Figure 35: Global earthquakes causal relationship for target cluster35, highlighted in green, all other driving areas highlighted in red

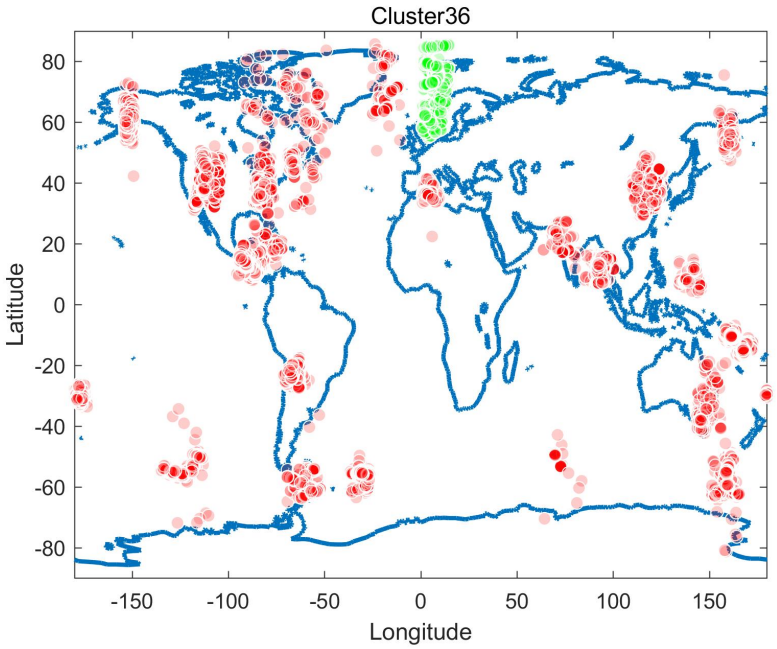


Figure 36: Global earthquakes causal relationship for target cluster36, highlighted in green, all other driving areas highlighted in red

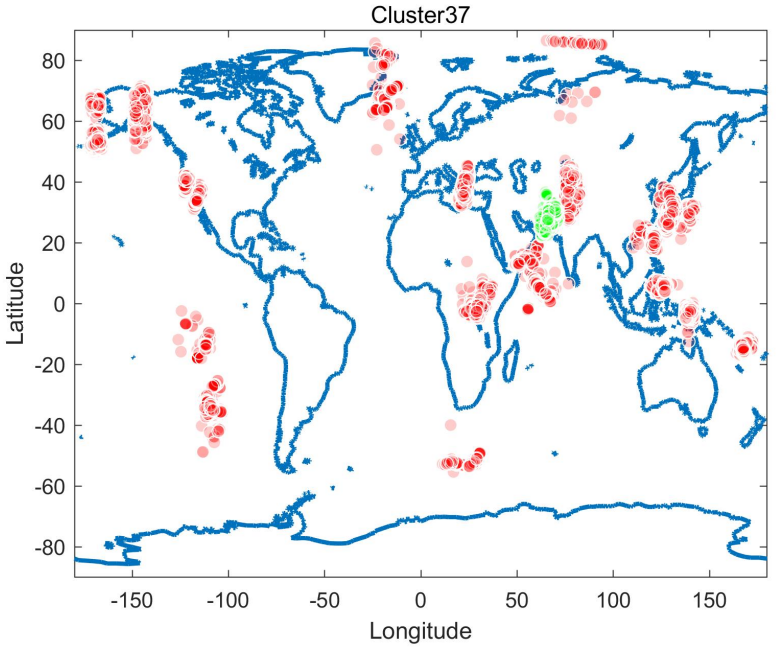


Figure 37: Global earthquakes causal relationship for target cluster37, highlighted in green, all other driving areas highlighted in red

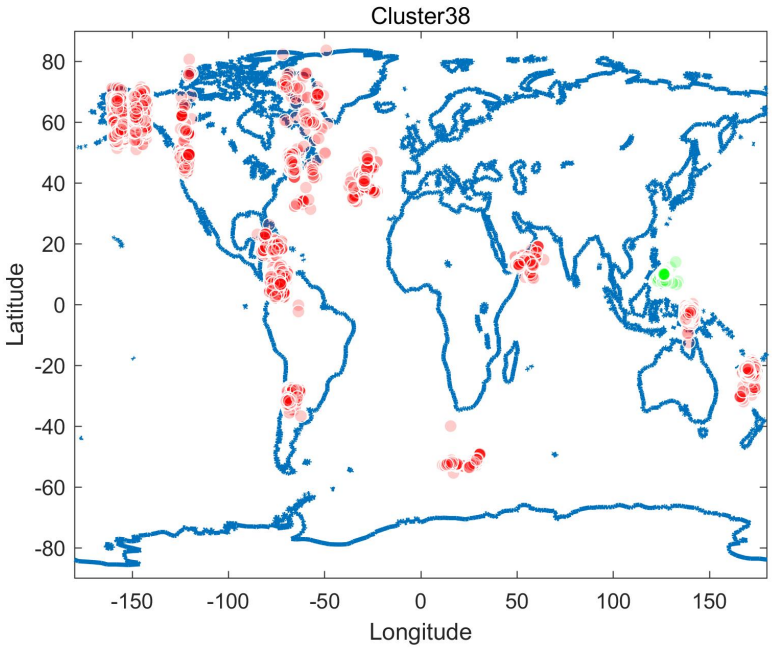


Figure 38: Global earthquakes causal relationship for target cluster38, highlighted in green, all other driving areas highlighted in red

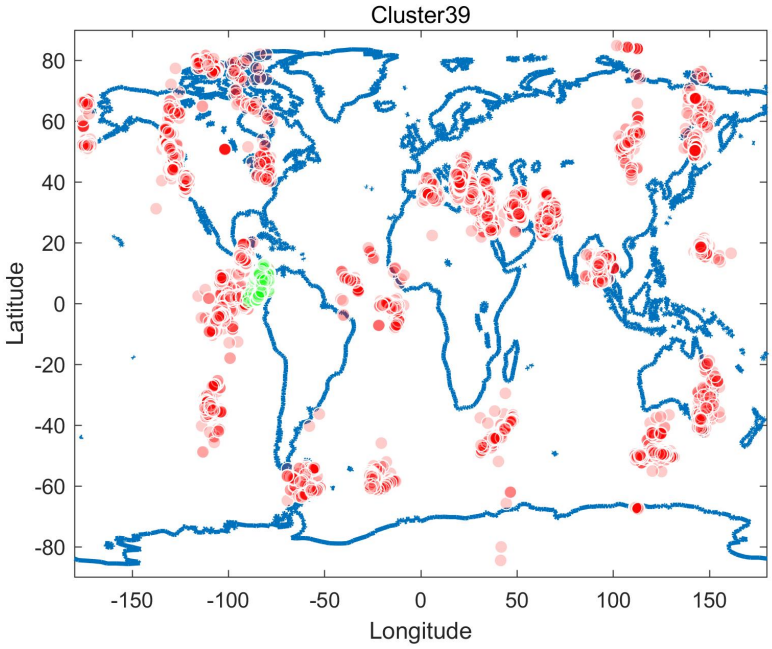


Figure 39: Global earthquakes causal relationship for target cluster39, highlighted in green, all other driving areas highlighted in red

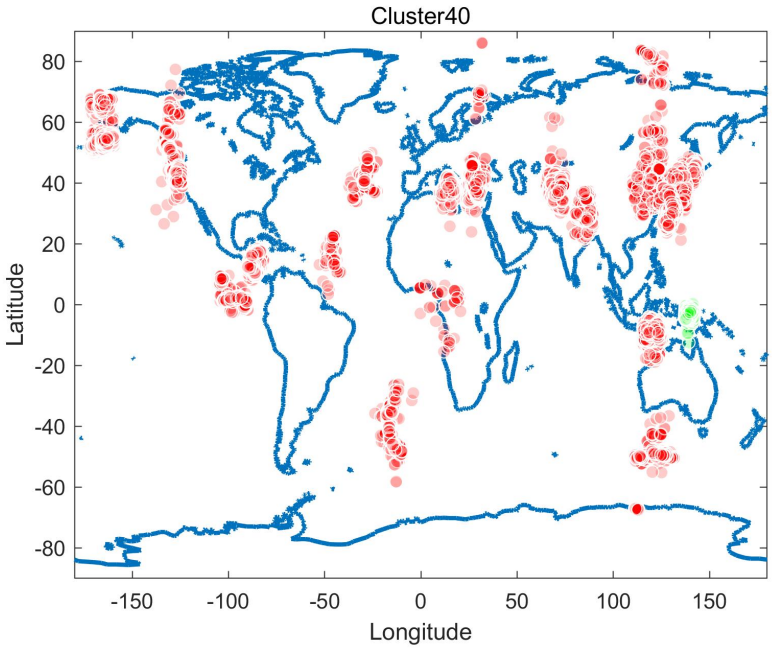


Figure 40: Global earthquakes causal relationship for target cluster40, highlighted in green, all other driving areas highlighted in red

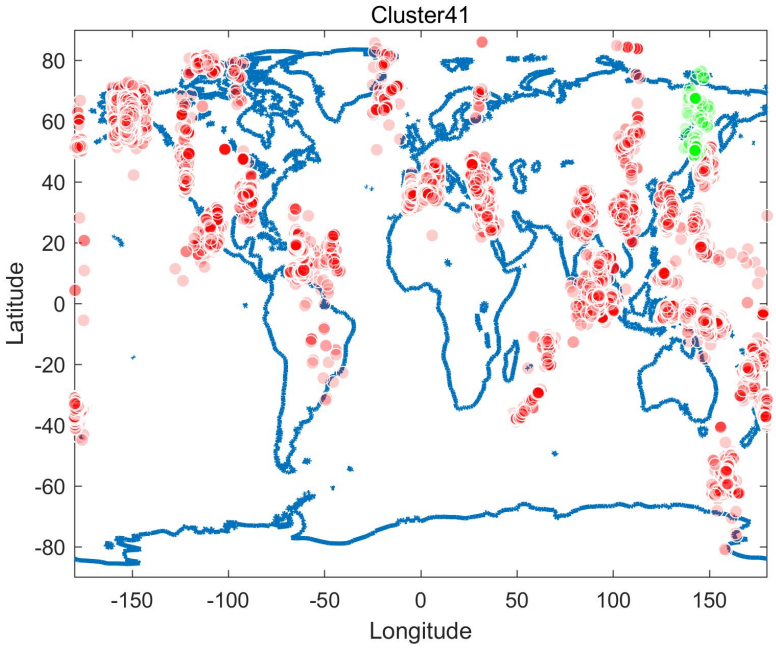


Figure 41: Global earthquakes causal relationship for target cluster41, highlighted in green, all other driving areas highlighted in red

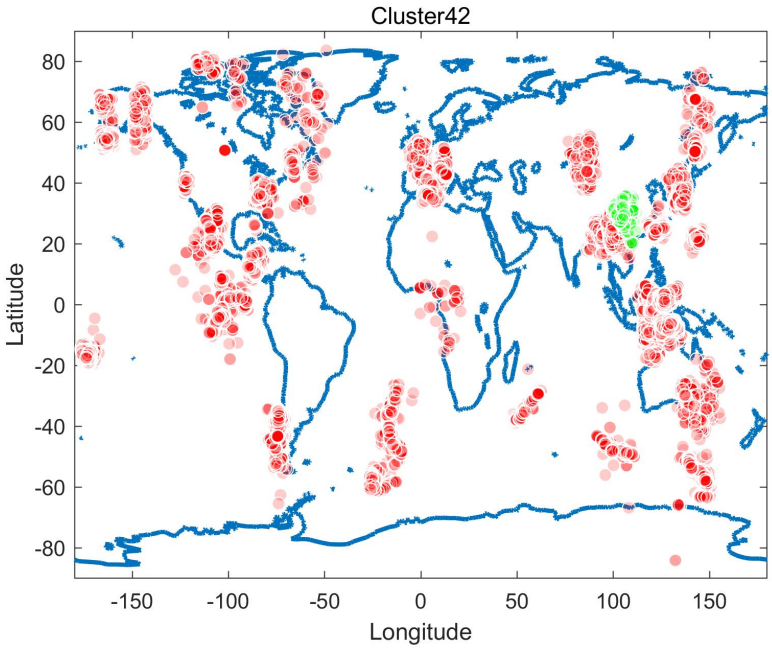


Figure 42: Global earthquakes causal relationship for target cluster42, highlighted in green, all other driving areas highlighted in red

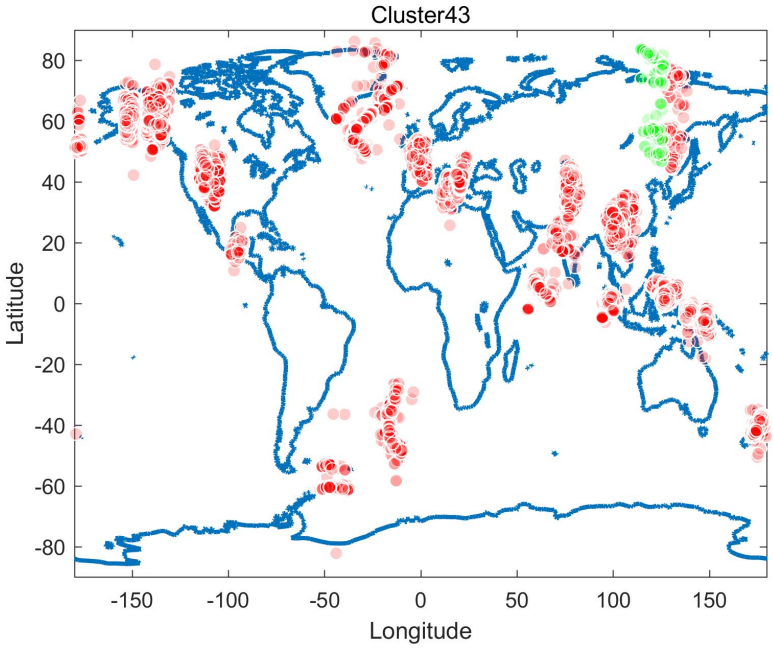


Figure 43: Global earthquakes causal relationship for target cluster43, highlighted in green, all other driving areas highlighted in red

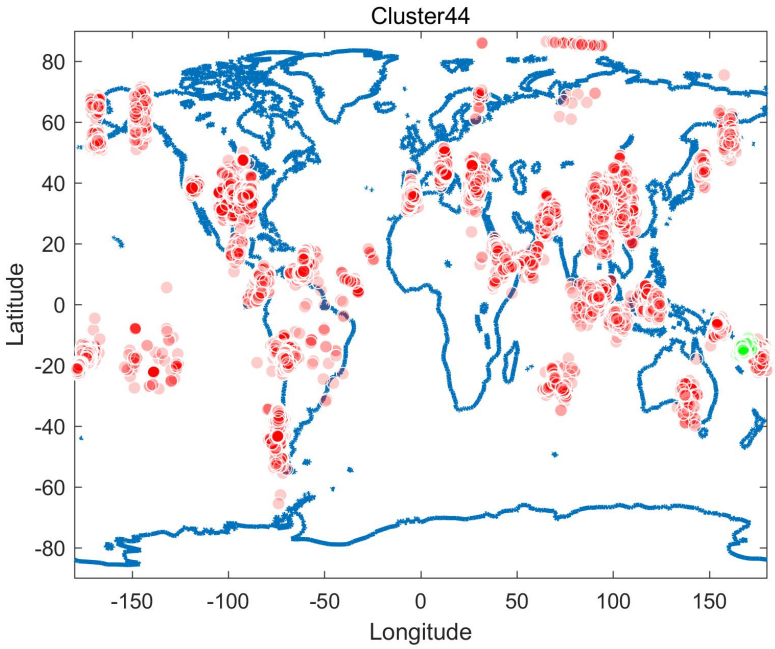


Figure 44: Global earthquakes causal relationship for target cluster44, highlighted in green, all other driving areas highlighted in red

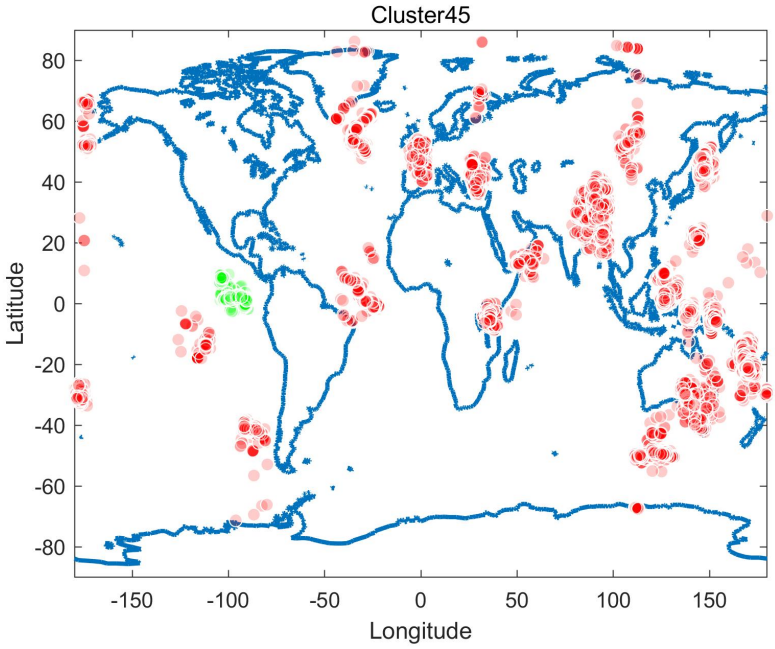


Figure 45: Global earthquakes causal relationship for target cluster45, highlighted in green, all other driving areas highlighted in red

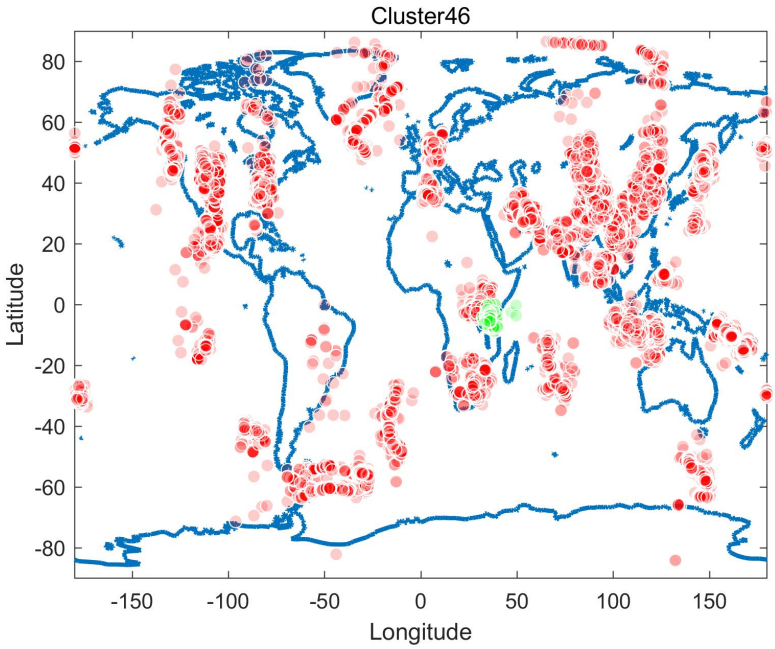


Figure 46: Global earthquakes causal relationship for target cluster46, highlighted in green, all other driving areas highlighted in red

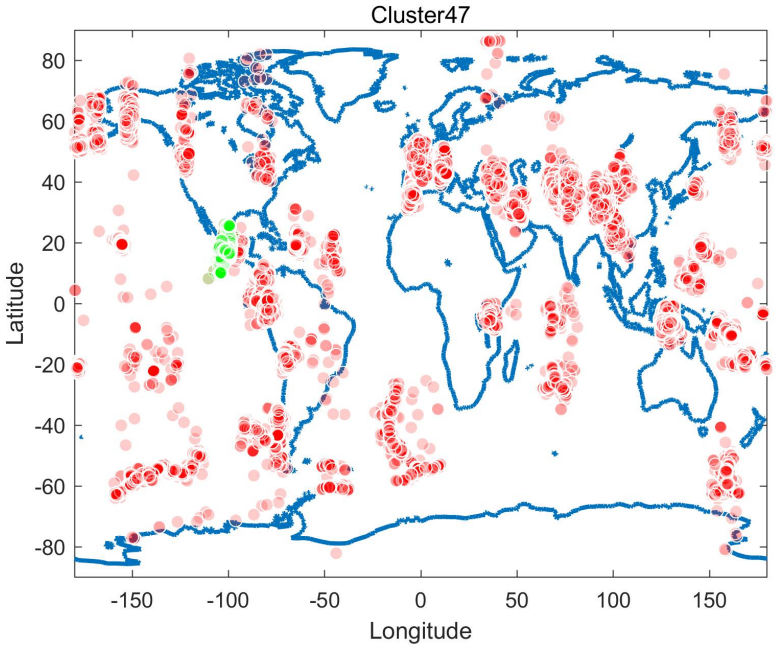


Figure 47: Global earthquakes causal relationship for target cluster47, highlighted in green, all other driving areas highlighted in red

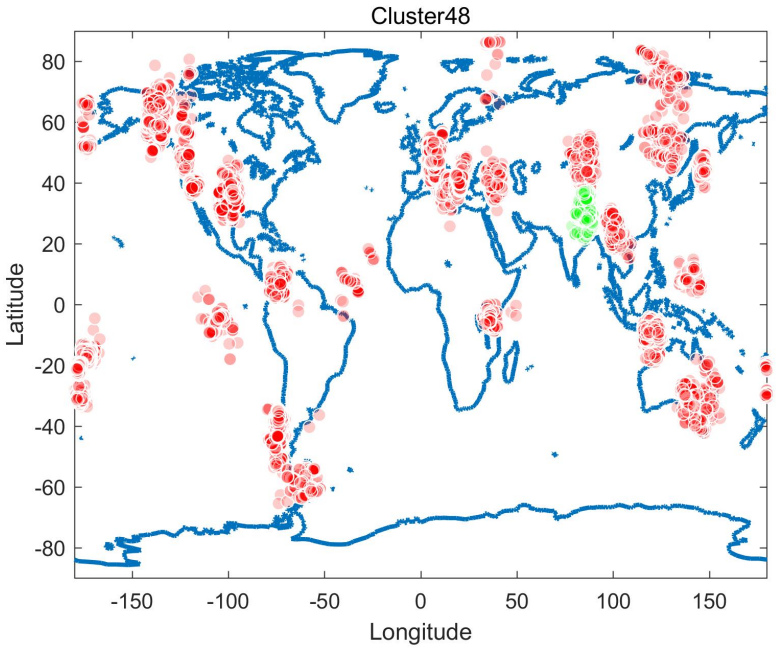


Figure 48: Global earthquakes causal relationship for target cluster48, highlighted in green, all other driving areas highlighted in red

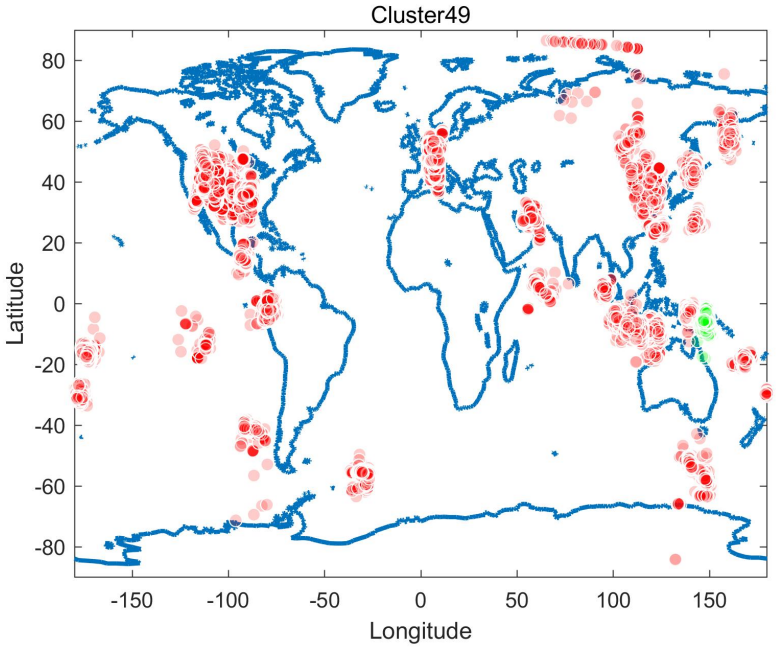


Figure 49: Global earthquakes causal relationship for target cluster49, highlighted in green, all other driving areas highlighted in red

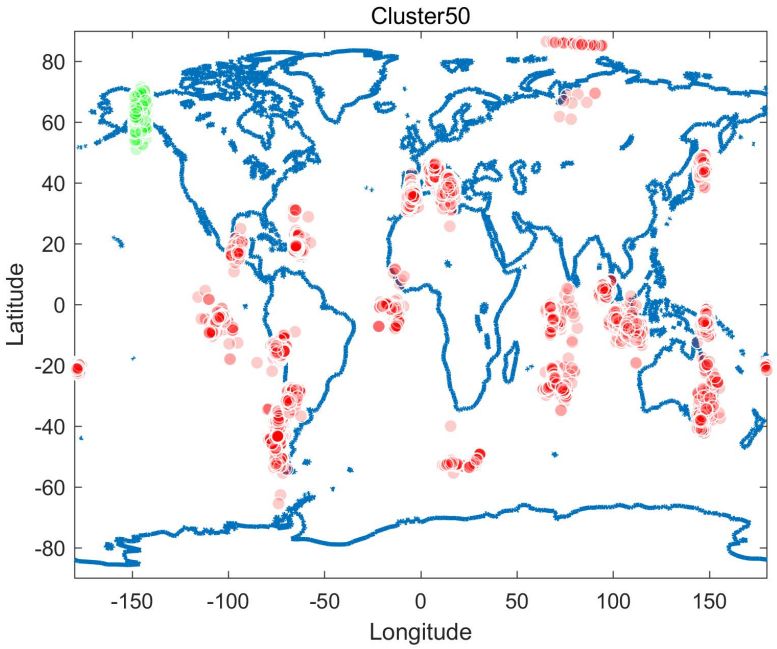


Figure 50: Global earthquakes causal relationship for target cluster50, highlighted in green, all other driving areas highlighted in red

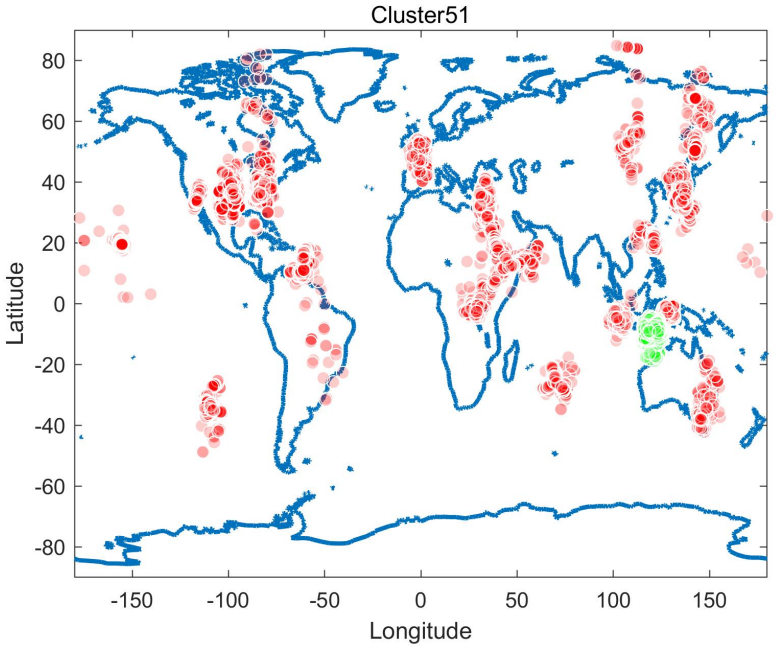


Figure 51: Global earthquakes causal relationship for target cluster51, highlighted in green, all other driving areas highlighted in red

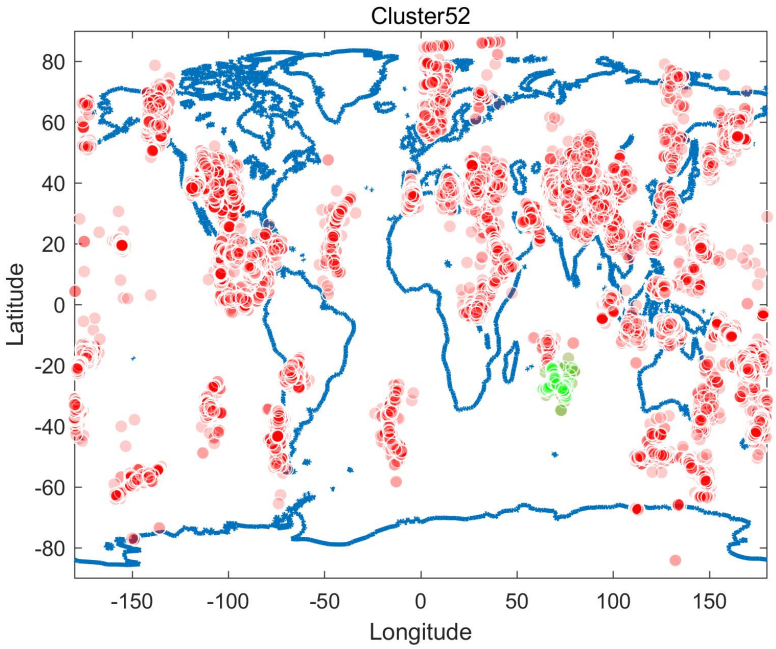


Figure 52: Global earthquakes causal relationship for target cluster52, highlighted in green, all other driving areas highlighted in red

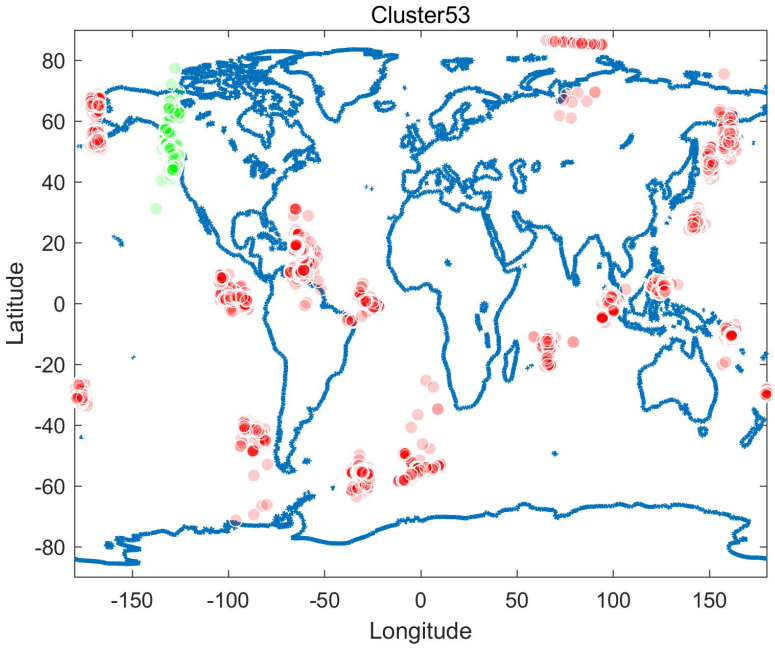


Figure 53: Global earthquakes causal relationship for target cluster53, highlighted in green, all other driving areas highlighted in red

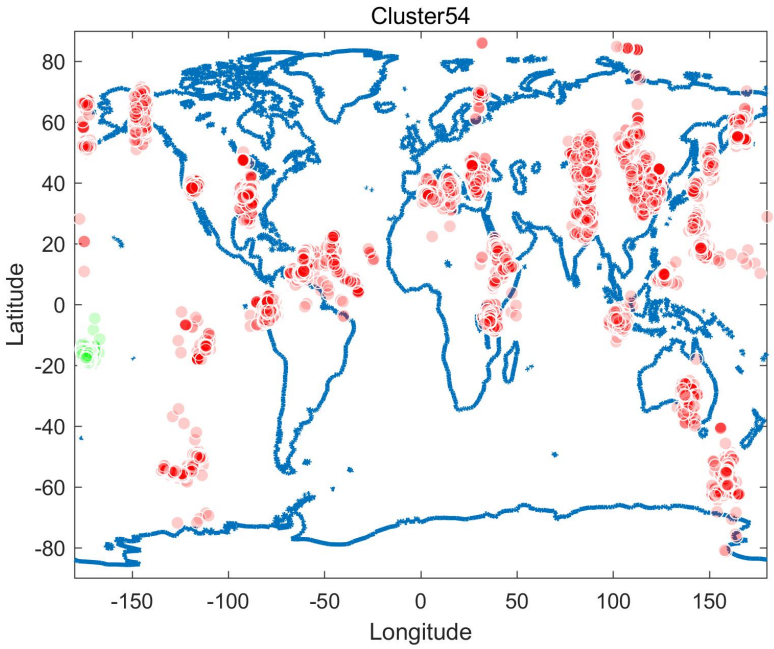


Figure 54: Global earthquakes causal relationship for target cluster54, highlighted in green, all other driving areas highlighted in red

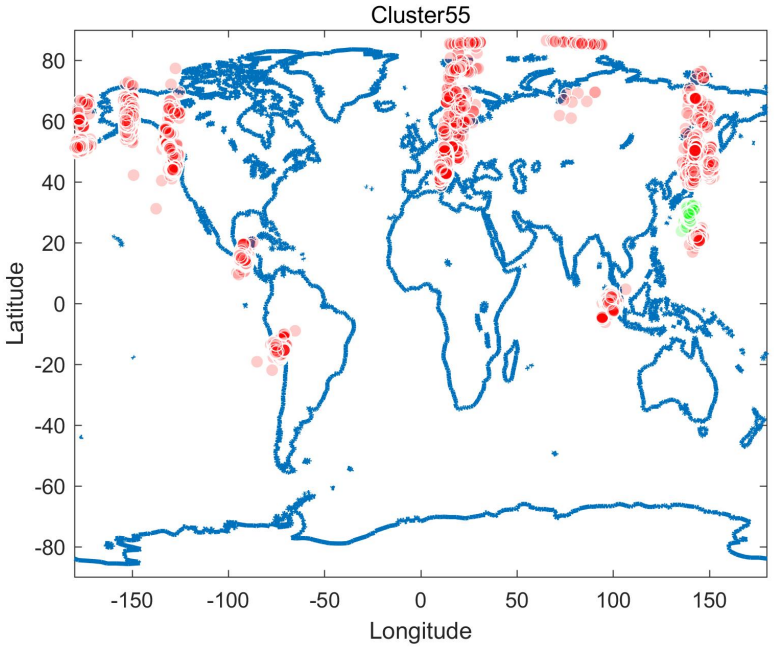


Figure 55: Global earthquakes causal relationship for target cluster55, highlighted in green, all other driving areas highlighted in red

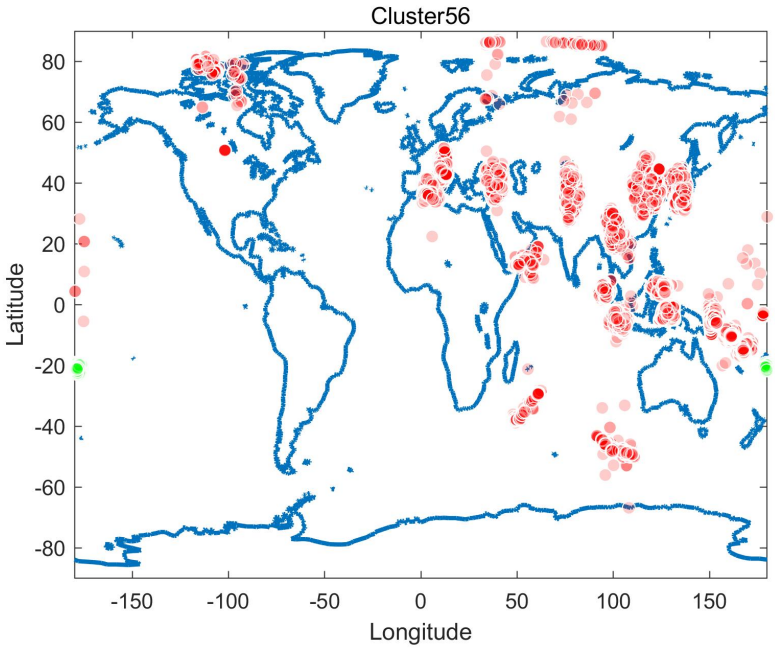


Figure 56: Global earthquakes causal relationship for target cluster56, highlighted in green, all other driving areas highlighted in red

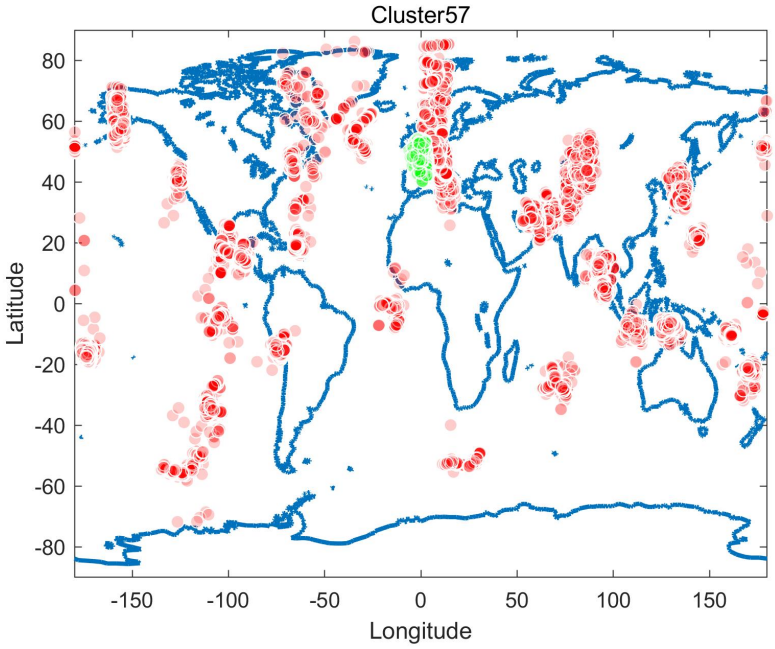


Figure 57: Global earthquakes causal relationship for target cluster57, highlighted in green, all other driving areas highlighted in red

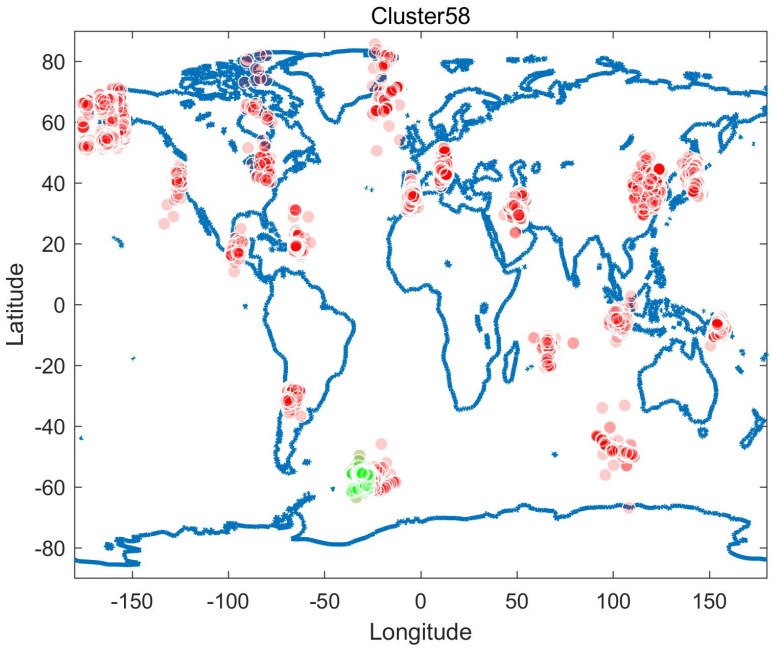


Figure 58: Global earthquakes causal relationship for target cluster58, highlighted in green, all other driving areas highlighted in red

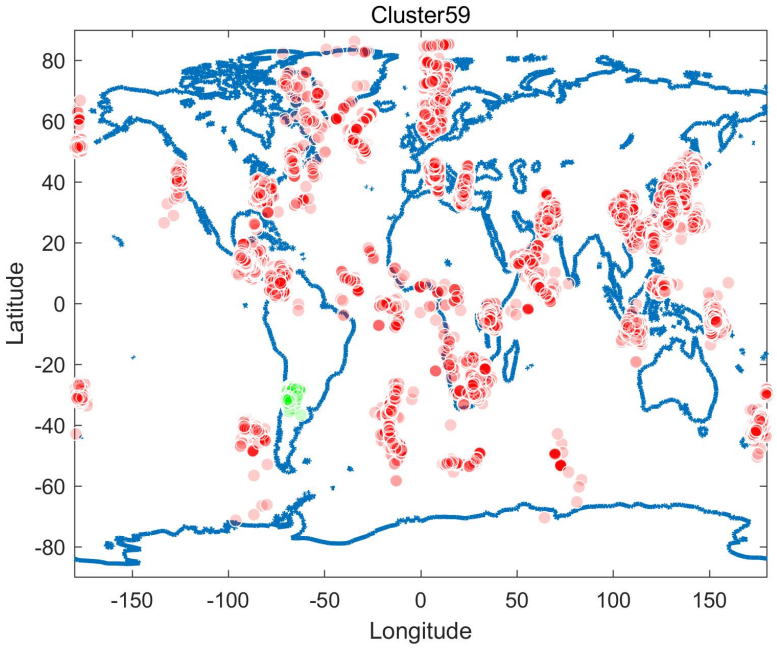


Figure 59: Global earthquakes causal relationship for target cluster59, highlighted in green, all other driving areas highlighted in red

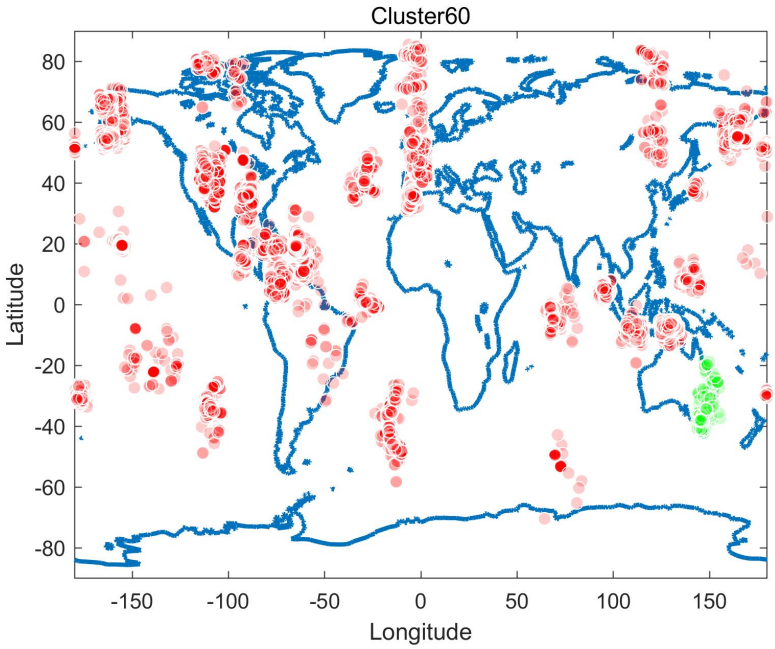


Figure 60: Global earthquakes causal relationship for target cluster60, highlighted in green, all other driving areas highlighted in red

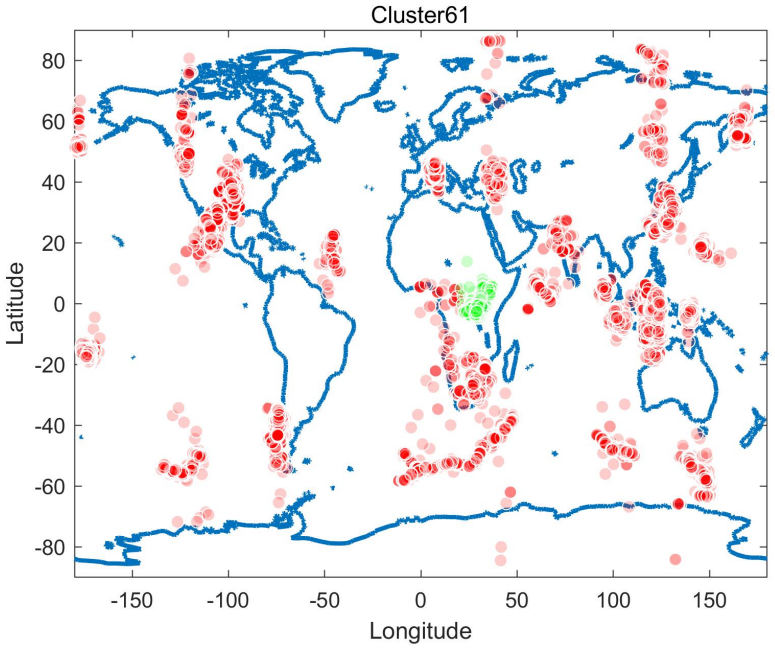


Figure 61: Global earthquakes causal relationship for target cluster61, highlighted in green, all other driving areas highlighted in red

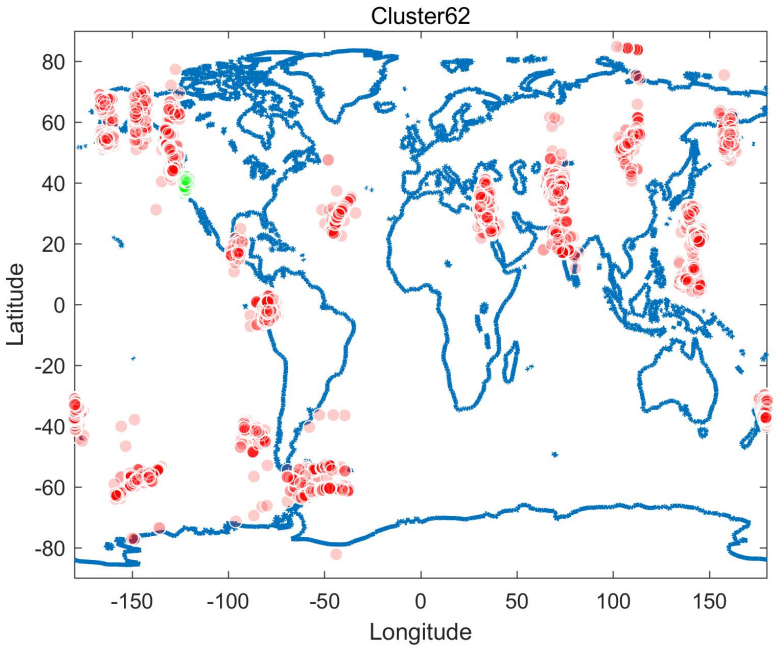


Figure 62: Global earthquakes causal relationship for target cluster62, highlighted in green, all other driving areas highlighted in red

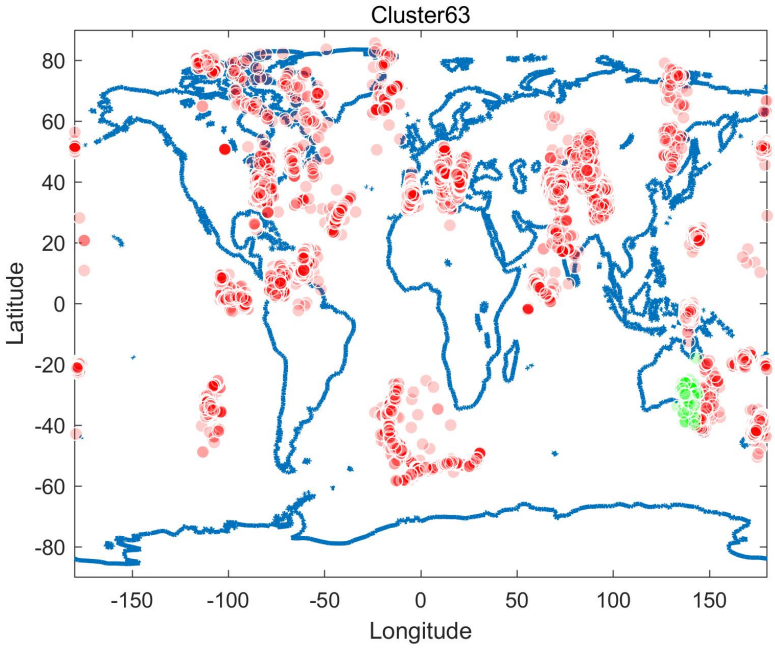


Figure 63: Global earthquakes causal relationship for target cluster63, highlighted in green, all other driving areas highlighted in red

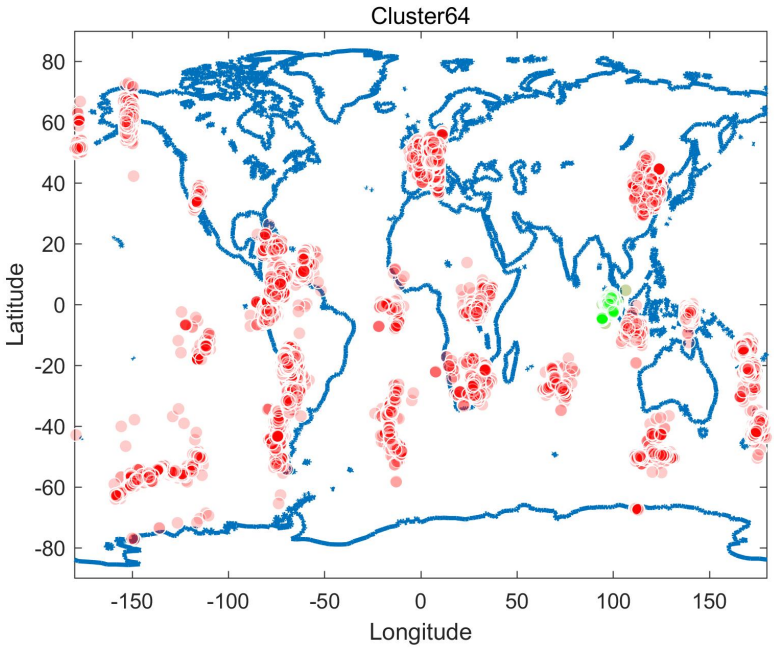


Figure 64: Global earthquakes causal relationship for target cluster64, highlighted in green, all other driving areas highlighted in red

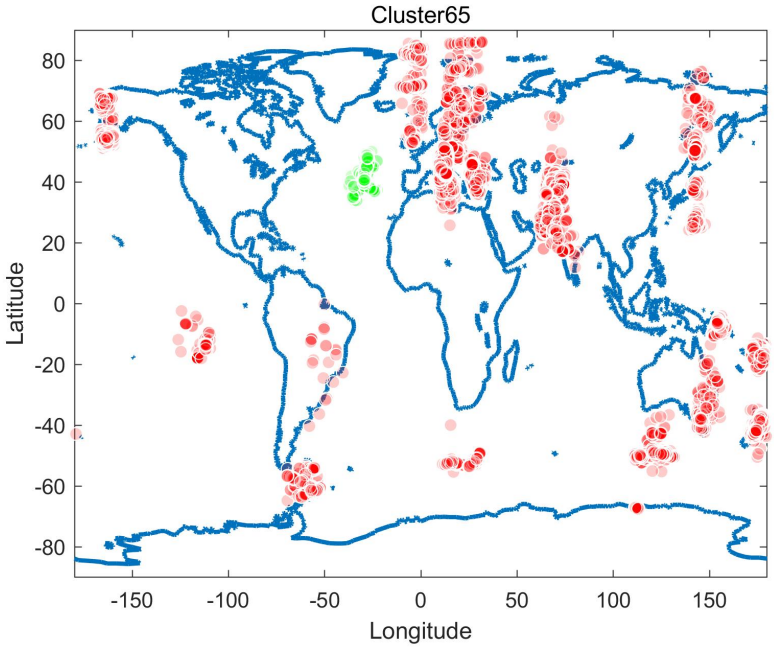


Figure 65: Global earthquakes causal relationship for target cluster65, highlighted in green, all other driving areas highlighted in red

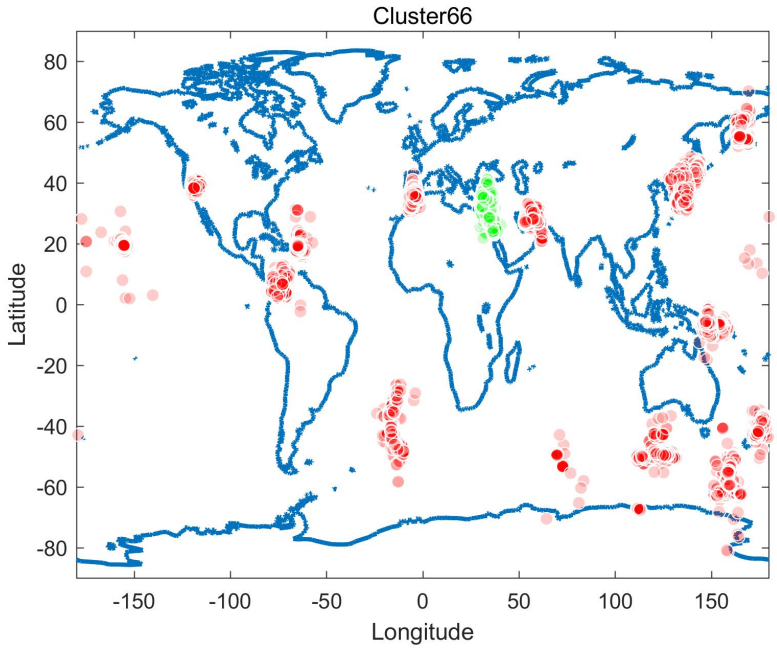


Figure 66: Global earthquakes causal relationship for target cluster66, highlighted in green, all other driving areas highlighted in red

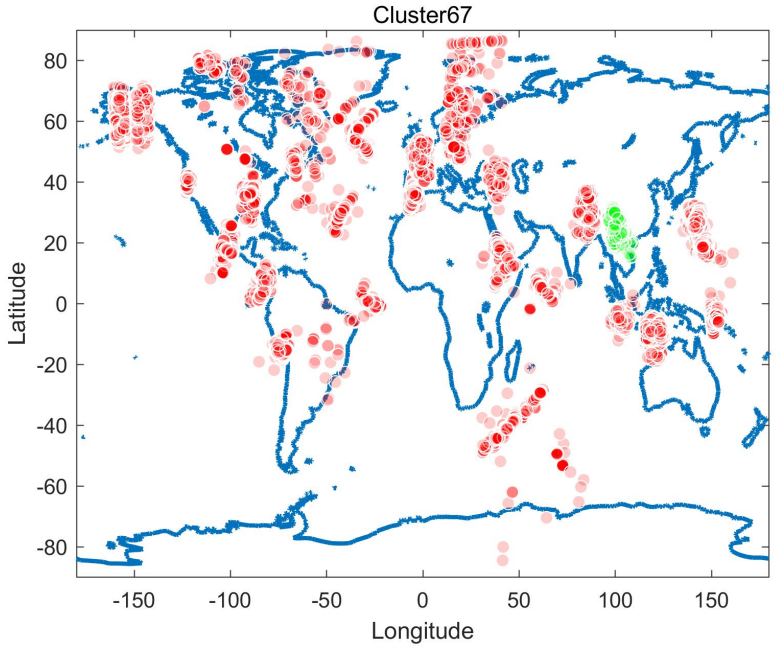


Figure 67: Global earthquakes causal relationship for target cluster67, highlighted in green, all other driving areas highlighted in red

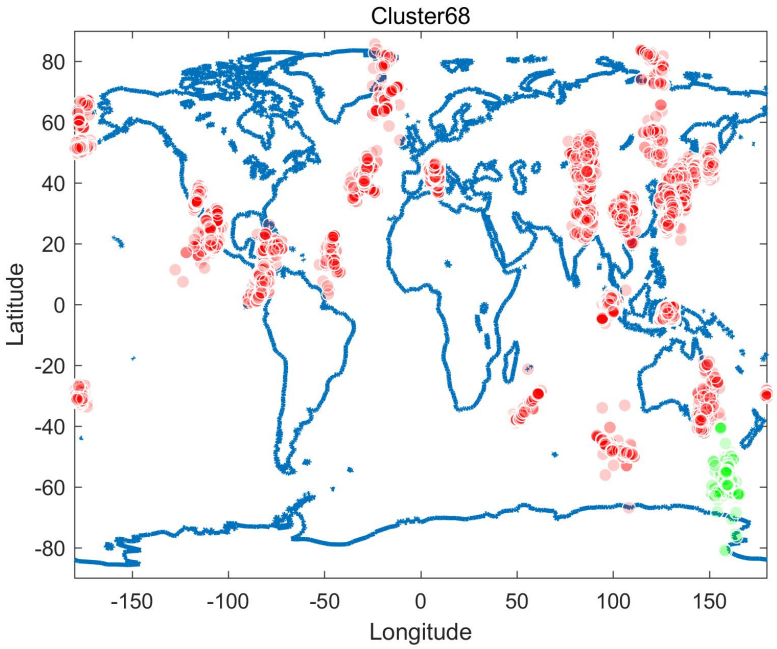


Figure 68: Global earthquakes causal relationship for target cluster68, highlighted in green, all other driving areas highlighted in red

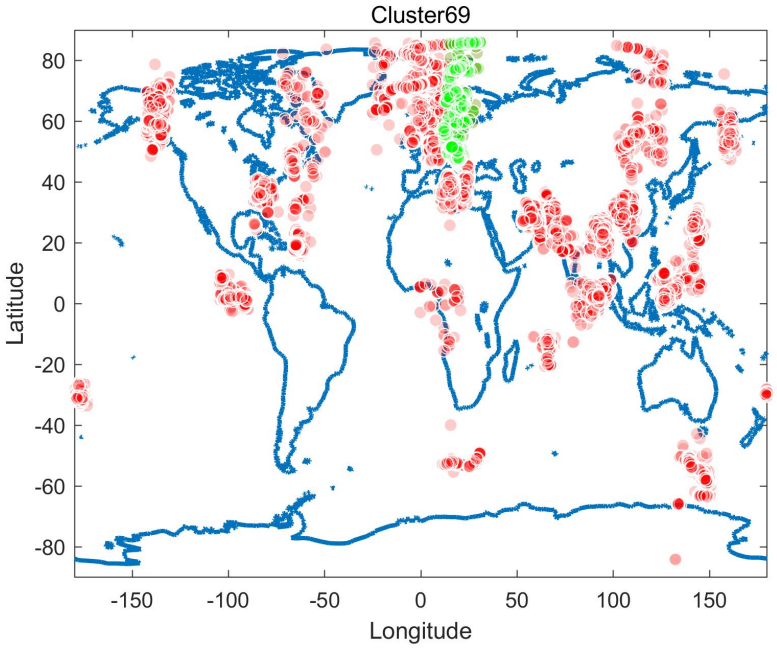


Figure 69: Global earthquakes causal relationship for target cluster69, highlighted in green, all other driving areas highlighted in red

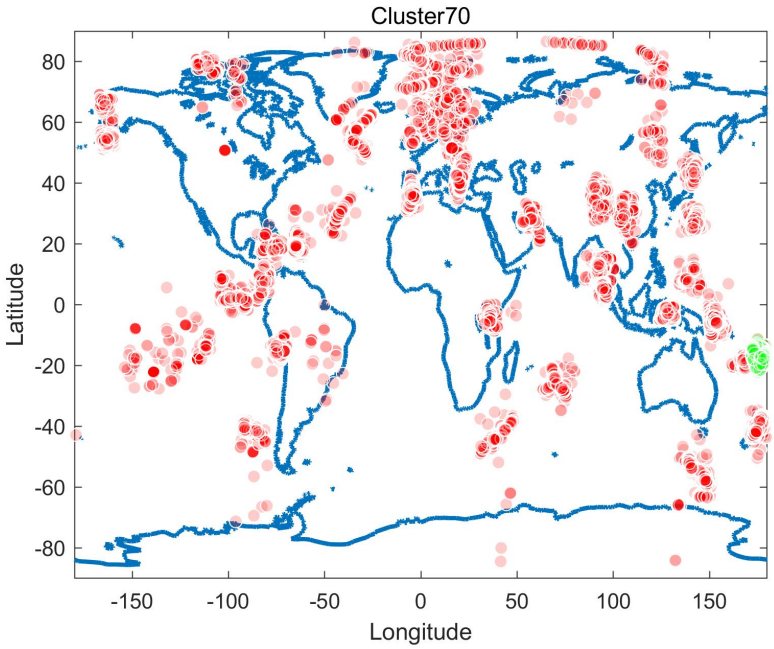


Figure 70: Global earthquakes causal relationship for target cluster70, highlighted in green, all other driving areas highlighted in red

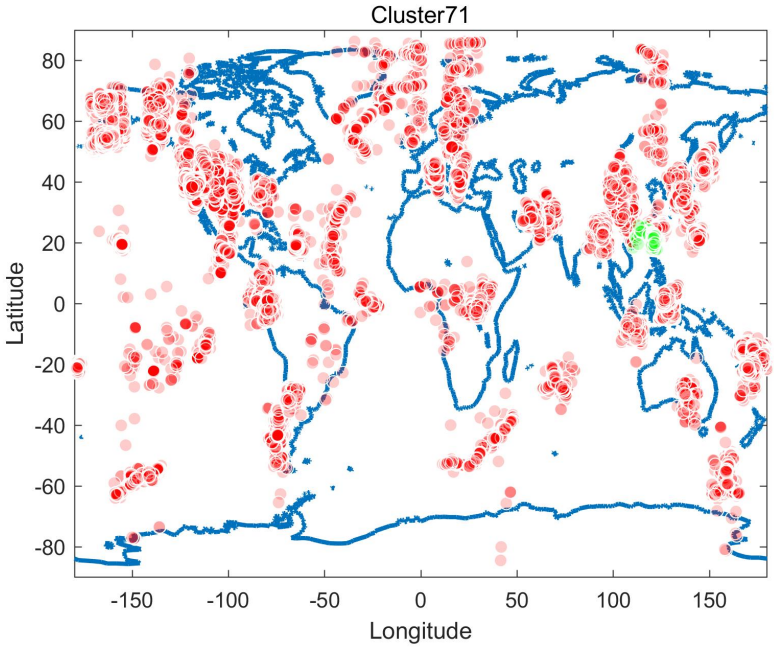


Figure 71: Global earthquakes causal relationship for target cluster71, highlighted in green, all other driving areas highlighted in red

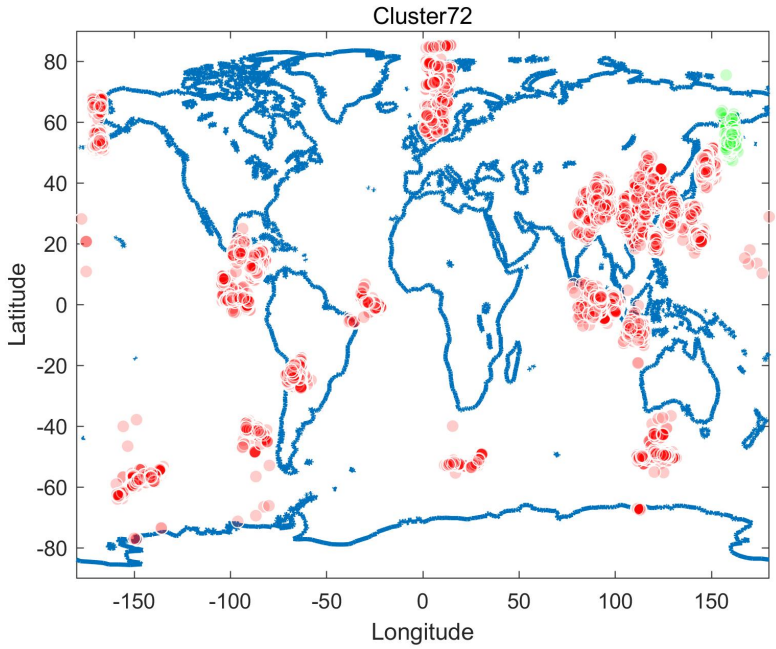


Figure 72: Global earthquakes causal relationship for target cluster72, highlighted in green, all other driving areas highlighted in red

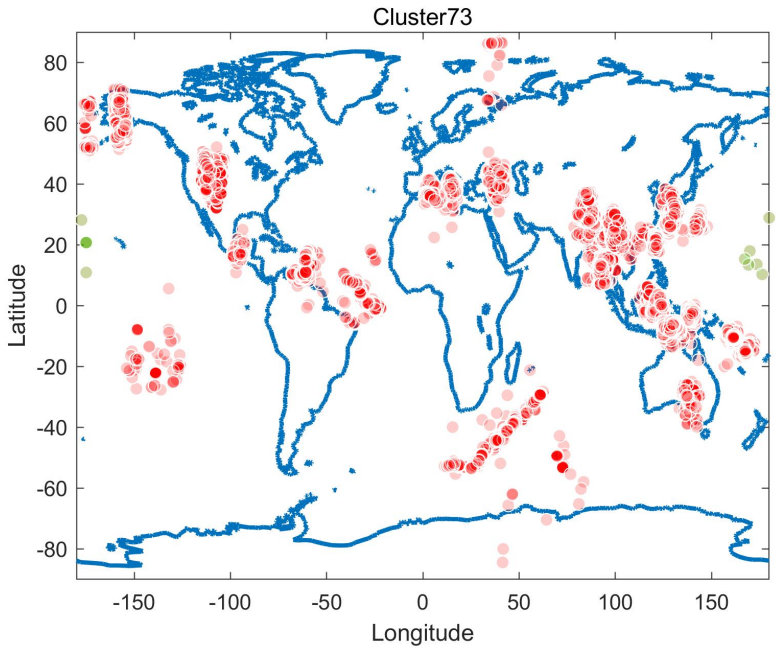


Figure 73: Global earthquakes causal relationship for target cluster73, highlighted in green, all other driving areas highlighted in red

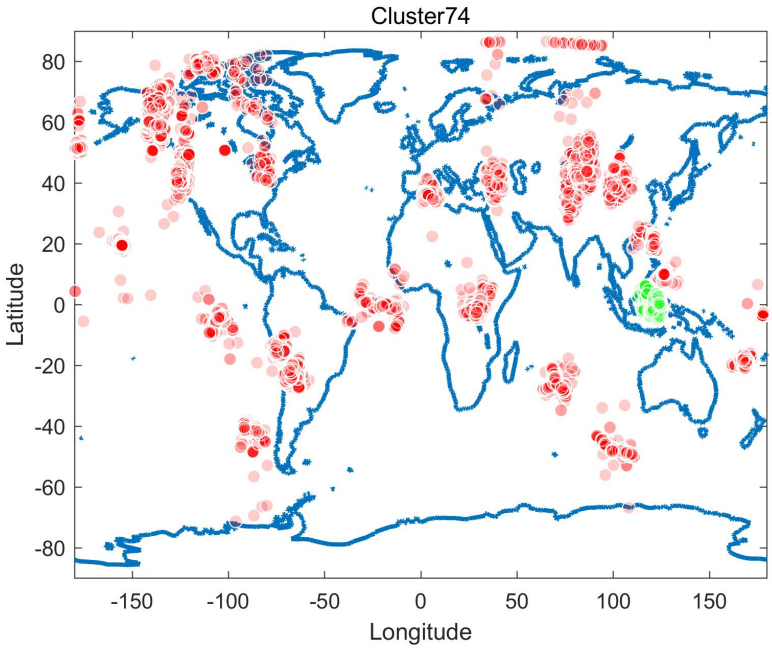


Figure 74: Global earthquakes causal relationship for target cluster74, highlighted in green, all other driving areas highlighted in red

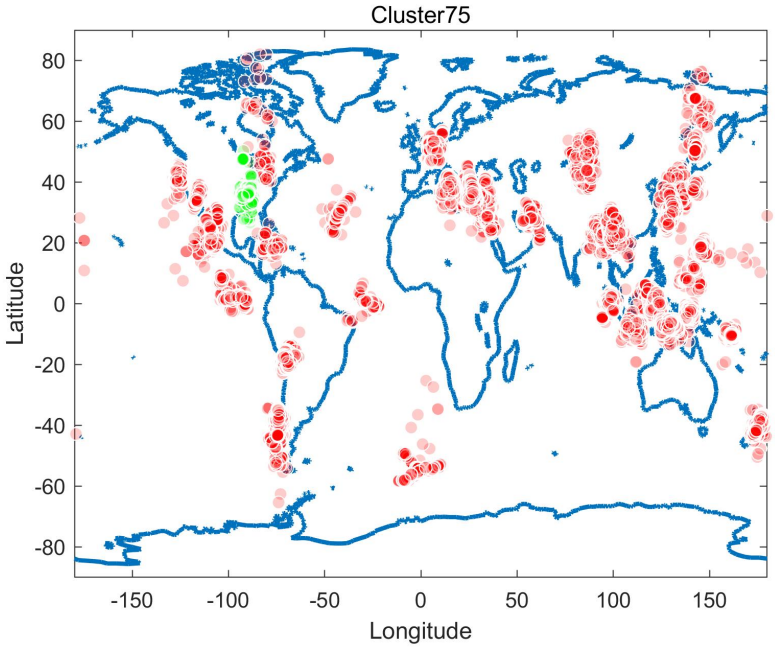


Figure 75: Global earthquakes causal relationship for target cluster75, highlighted in green, all other driving areas highlighted in red

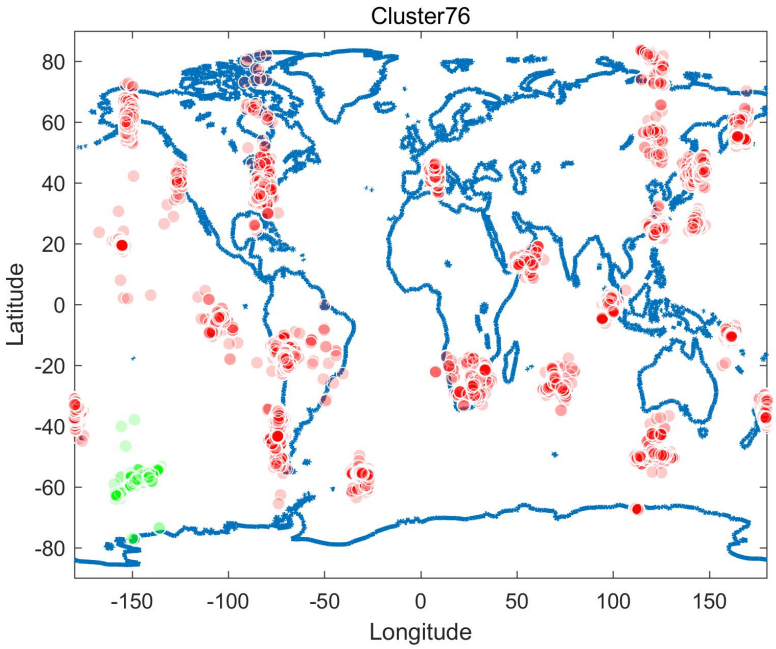


Figure 76: Global earthquakes causal relationship for target cluster76, highlighted in green, all other driving areas highlighted in red

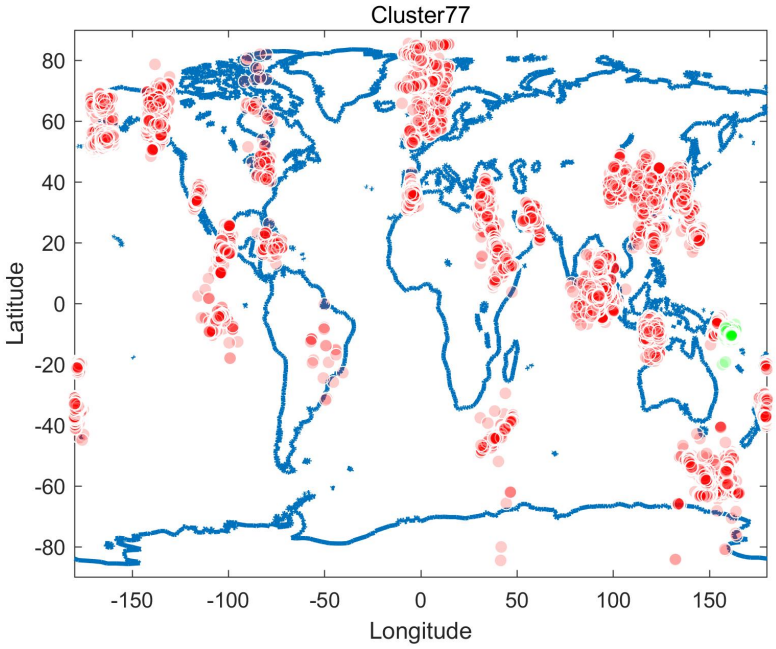


Figure 77: Global earthquakes causal relationship for target cluster77, highlighted in green, all other driving areas highlighted in red

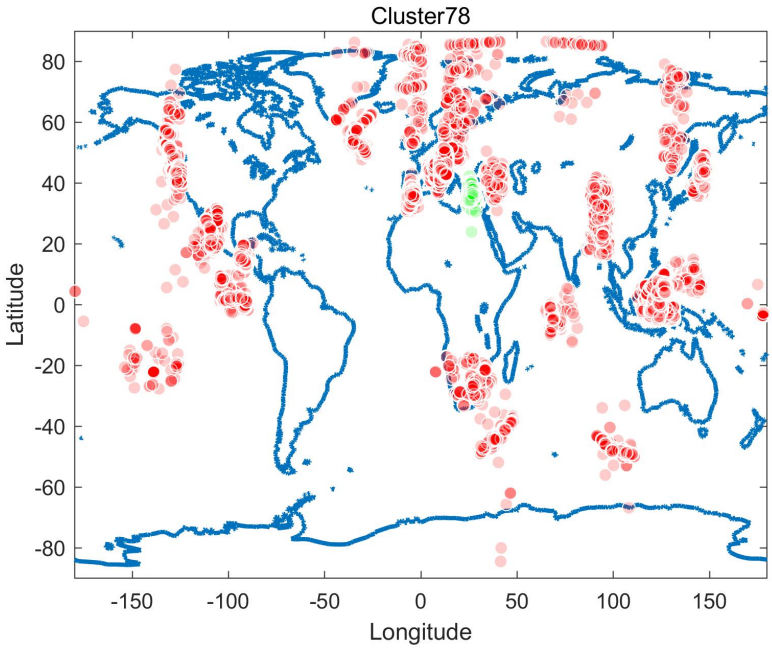


Figure 78: Global earthquakes causal relationship for target cluster78, highlighted in green, all other driving areas highlighted in red

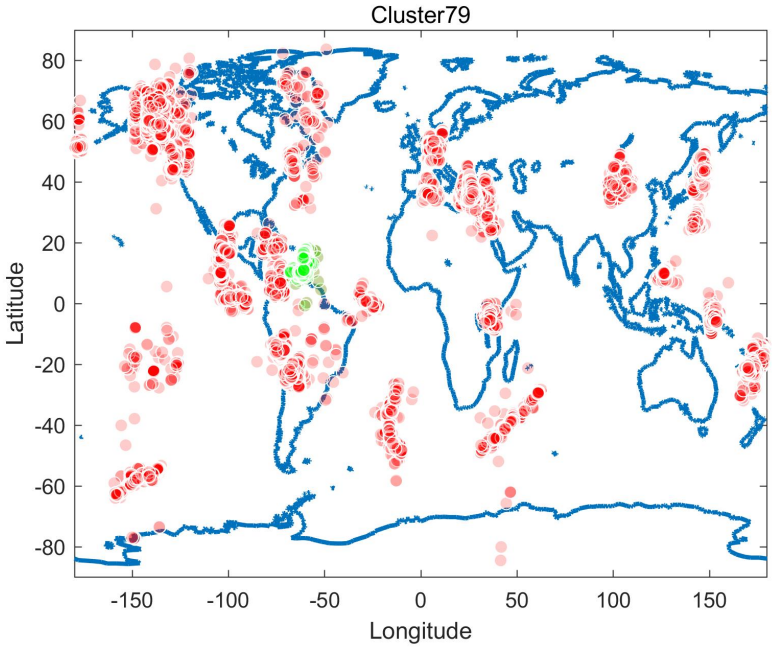


Figure 79: Global earthquakes causal relationship for target cluster79, highlighted in green, all other driving areas highlighted in red

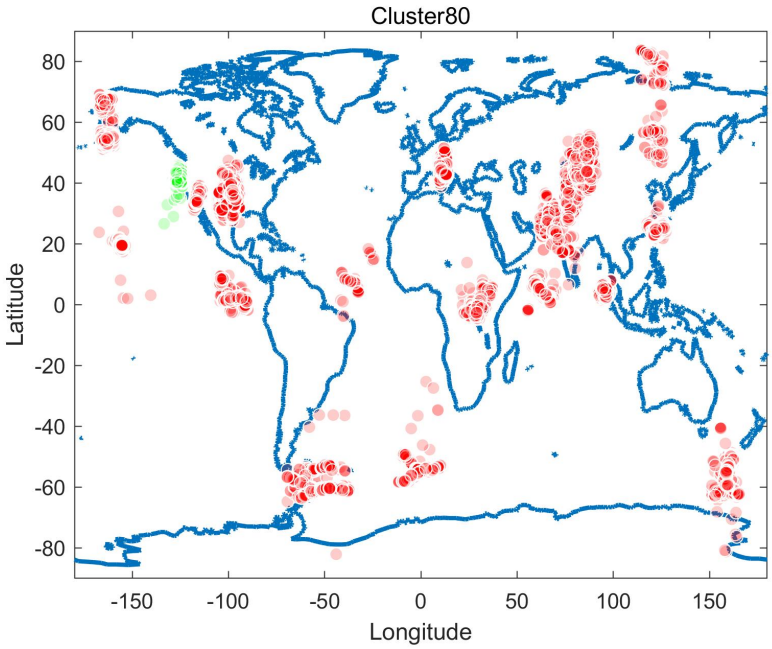


Figure 80: Global earthquakes causal relationship for target cluster80, highlighted in green, all other driving areas highlighted in red

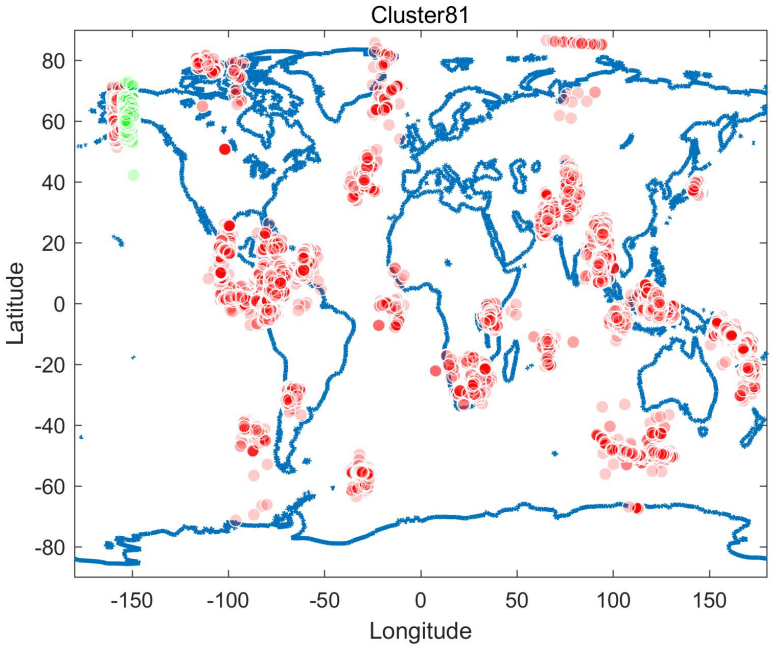


Figure 81: Global earthquakes causal relationship for target cluster81, highlighted in green, all other driving areas highlighted in red

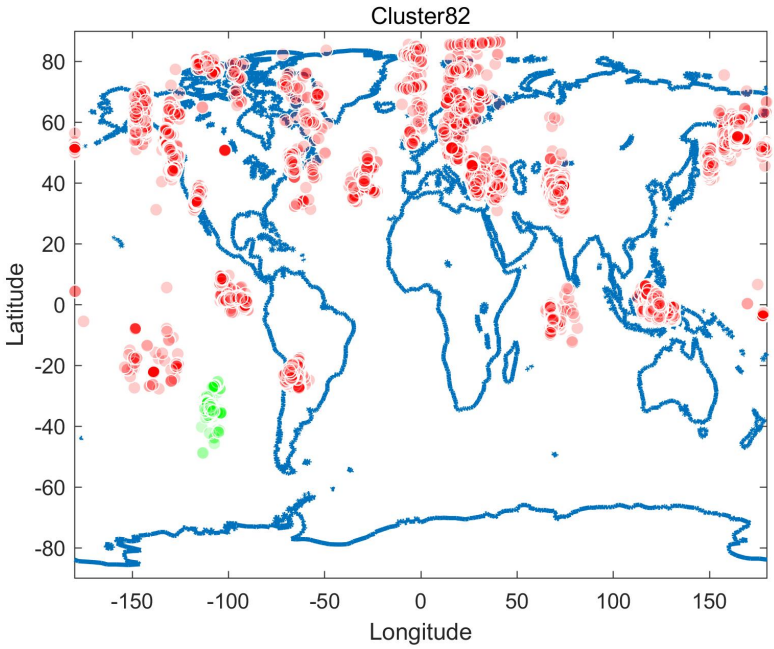


Figure 82: Global earthquakes causal relationship for target cluster82, highlighted in green, all other driving areas highlighted in red

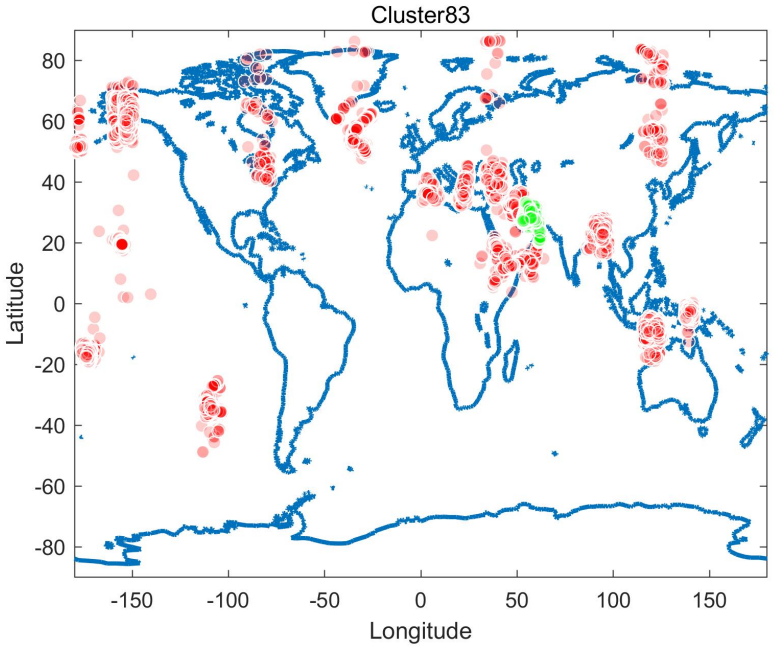


Figure 83: Global earthquakes causal relationship for target cluster83, highlighted in green, all other driving areas highlighted in red

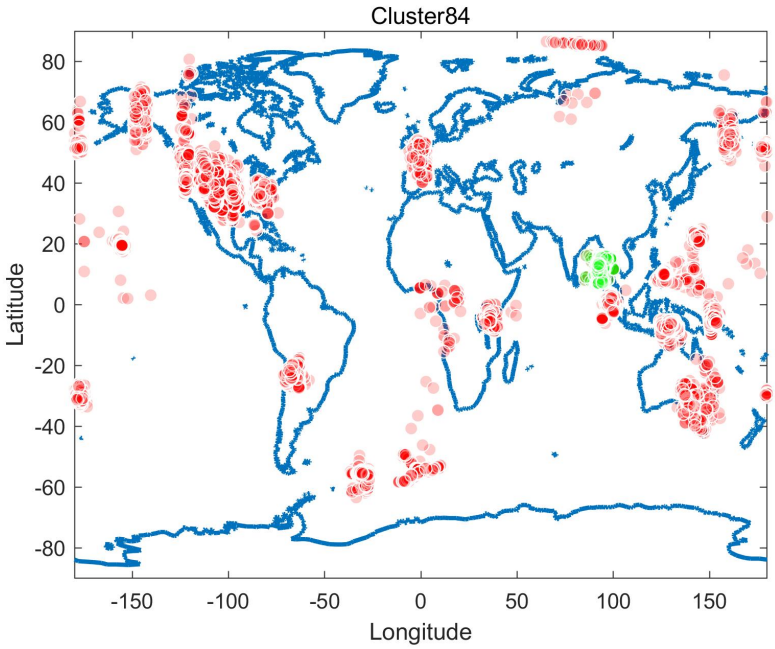


Figure 84: Global earthquakes causal relationship for target cluster84, highlighted in green, all other driving areas highlighted in red

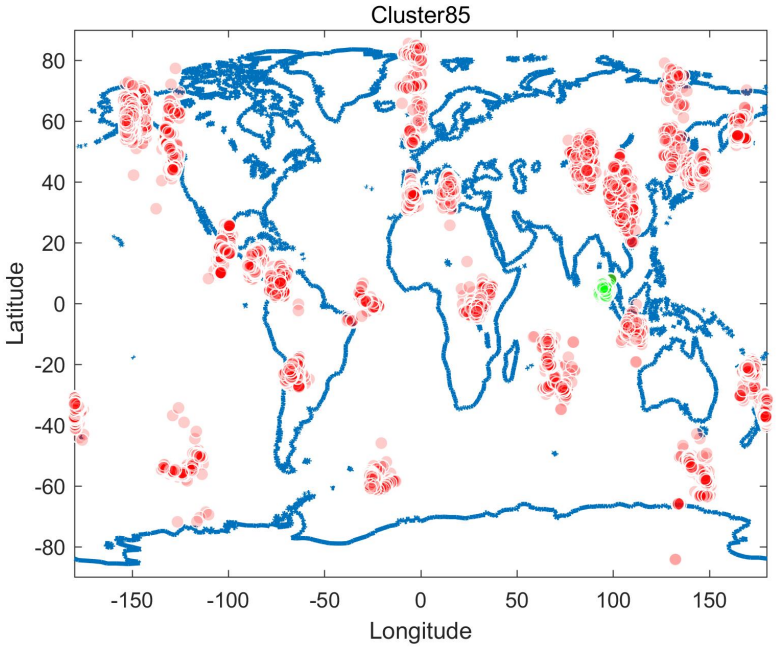


Figure 85: Global earthquakes causal relationship for target cluster85, highlighted in green, all other driving areas highlighted in red

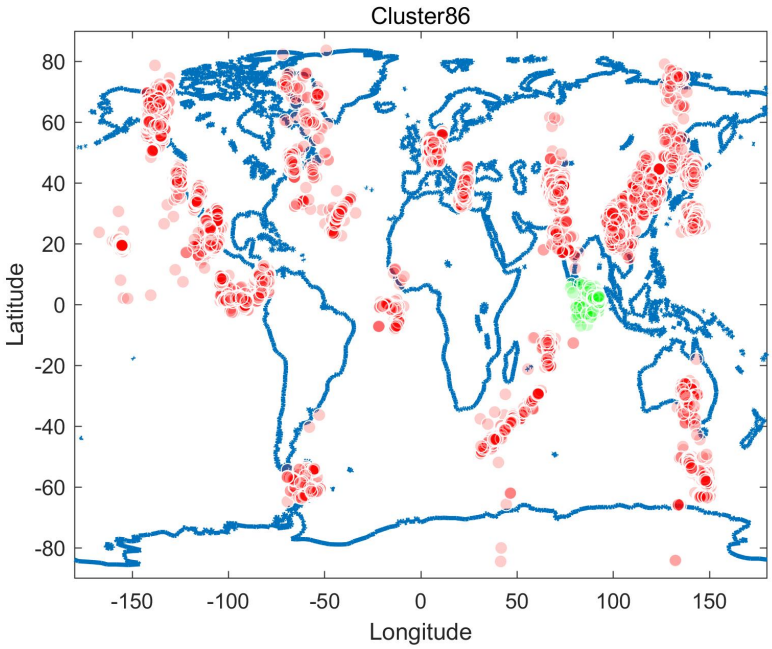


Figure 86: Global earthquakes causal relationship for target cluster86, highlighted in green, all other driving areas highlighted in red

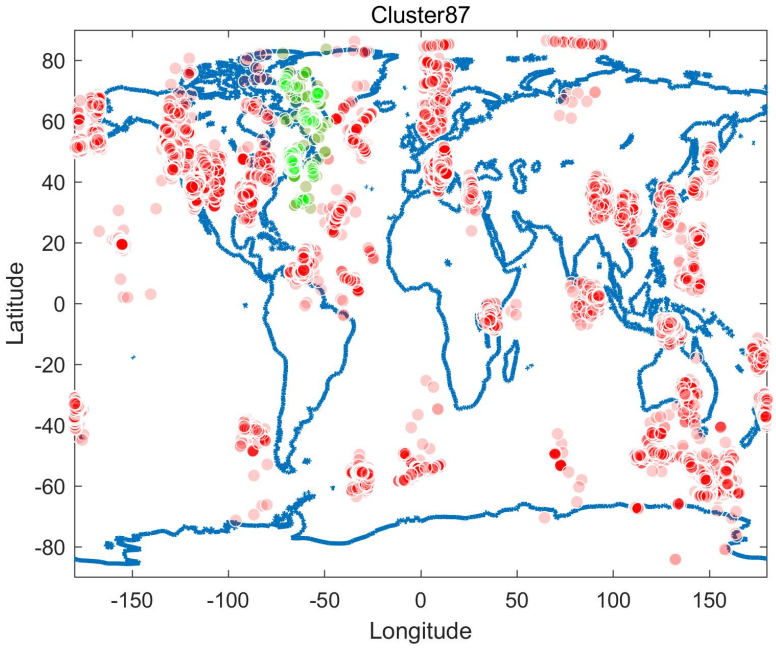


Figure 87: Global earthquakes causal relationship for target cluster87, highlighted in green, all other driving areas highlighted in red

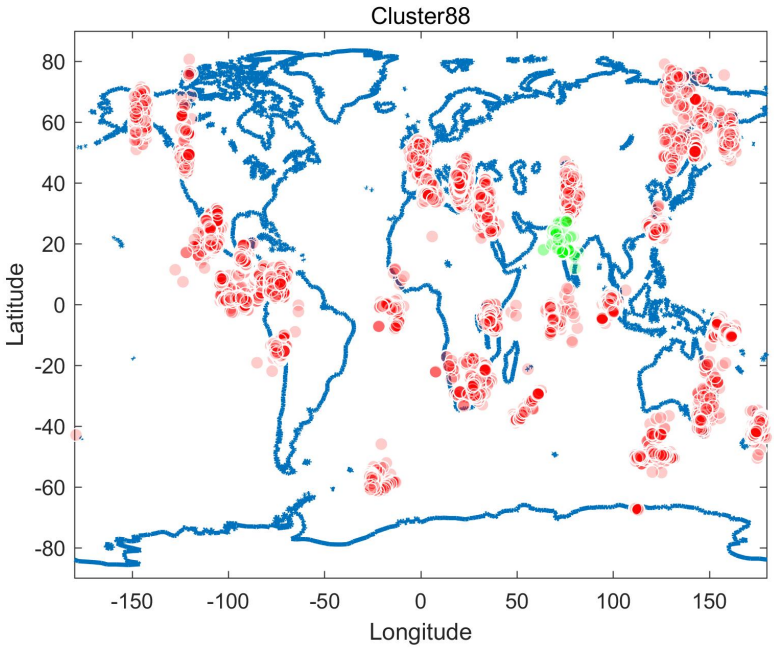


Figure 88: Global earthquakes causal relationship for target cluster88, highlighted in green, all other driving areas highlighted in red

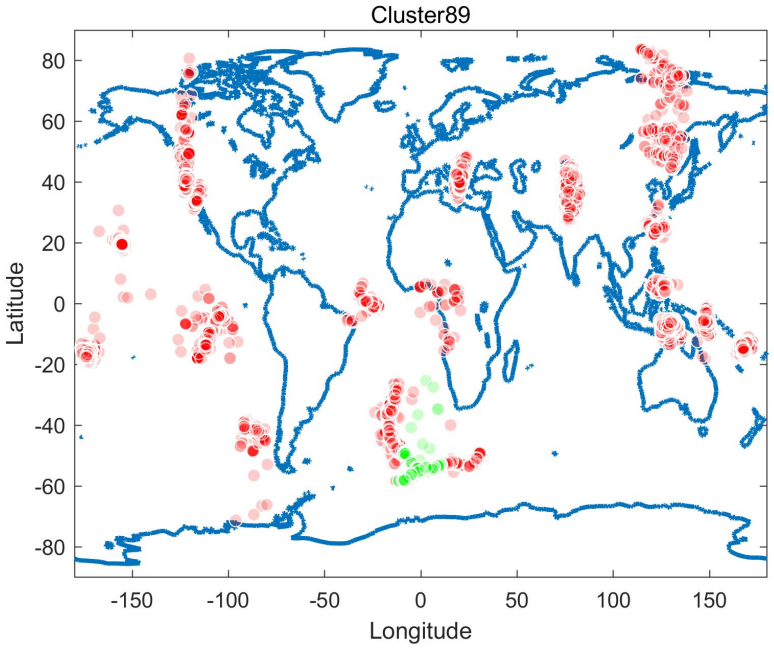


Figure 89: Global earthquakes causal relationship for target cluster89, highlighted in green, all other driving areas highlighted in red

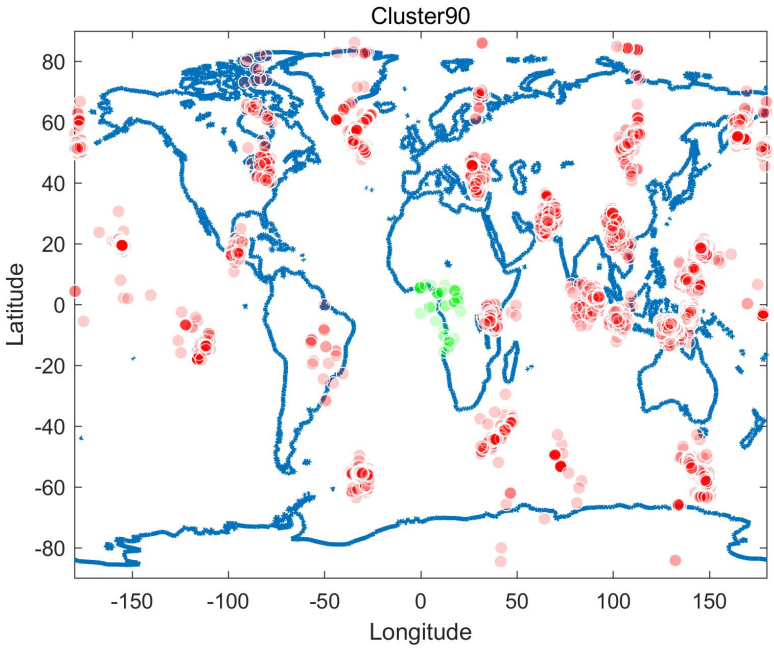


Figure 90: Global earthquakes causal relationship for target cluster90, highlighted in green, all other driving areas highlighted in red

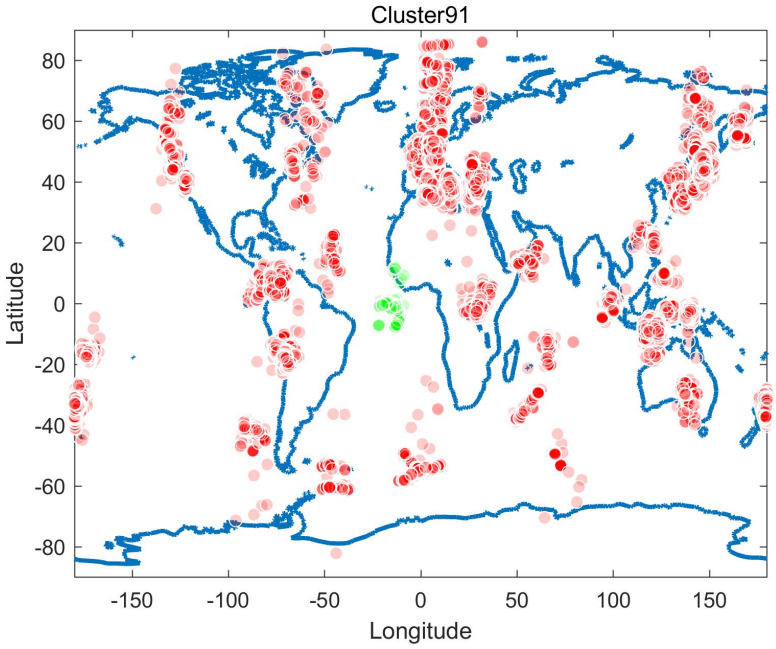


Figure 91: Global earthquakes causal relationship for target cluster91, highlighted in green, all other driving areas highlighted in red

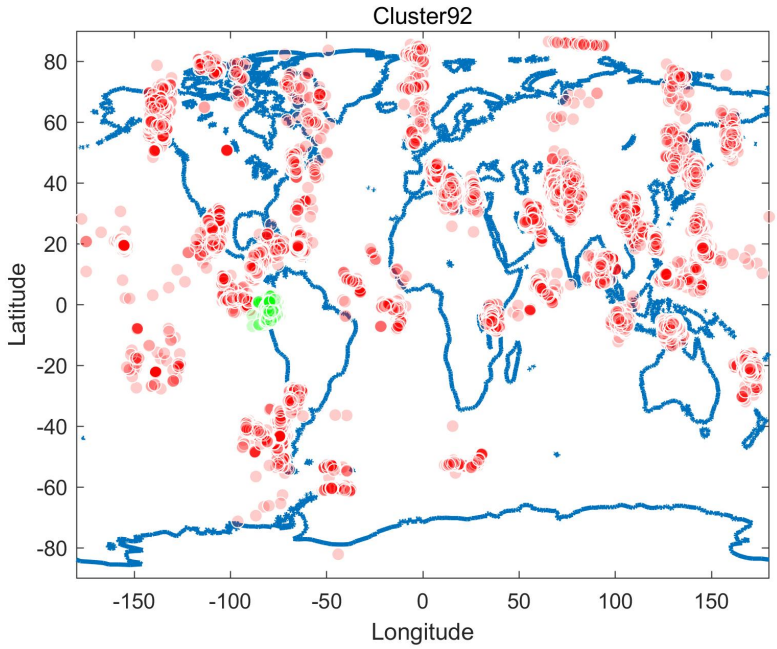


Figure 92: Global earthquakes causal relationship for target cluster92, highlighted in green, all other driving areas highlighted in red

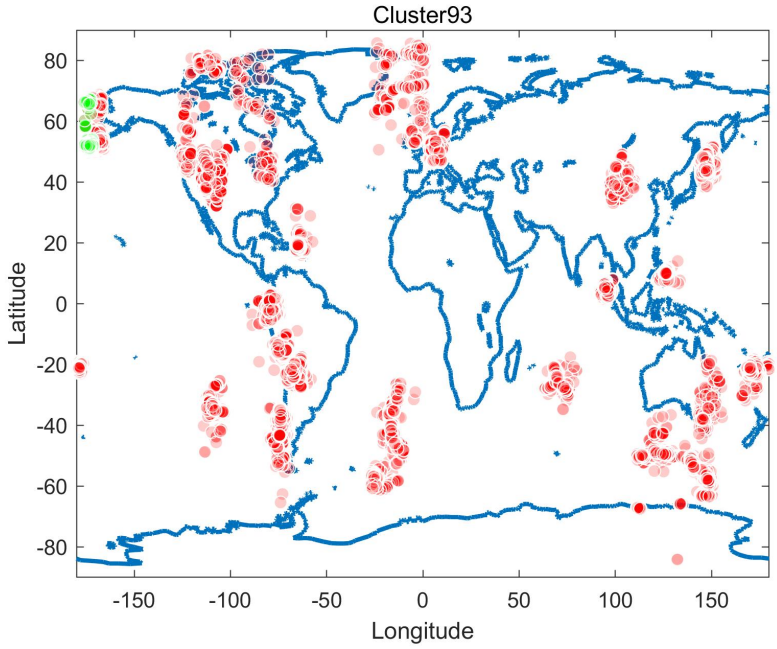


Figure 93: Global earthquakes causal relationship for target cluster93, highlighted in green, all other driving areas highlighted in red

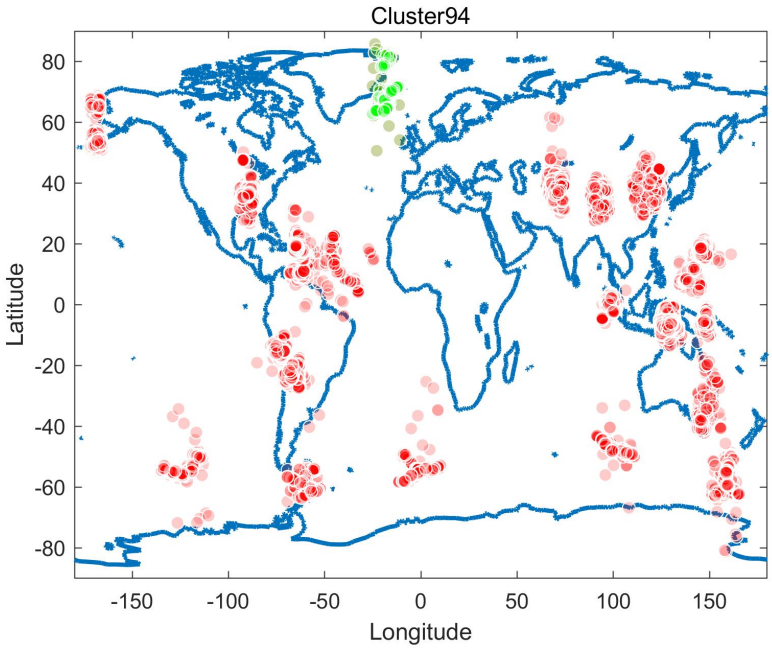


Figure 94: Global earthquakes causal relationship for target cluster94, highlighted in green, all other driving areas highlighted in red

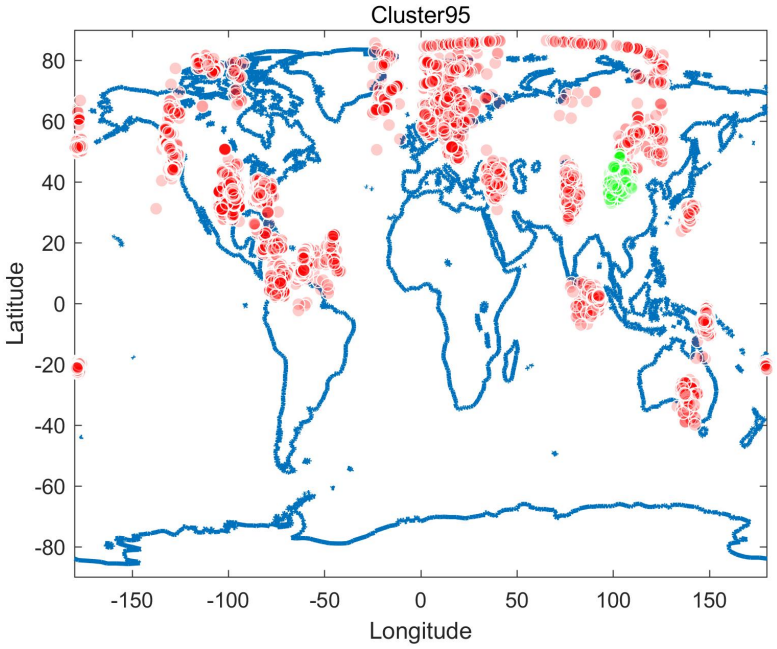


Figure 95: Global earthquakes causal relationship for target cluster95, highlighted in green, all other driving areas highlighted in red

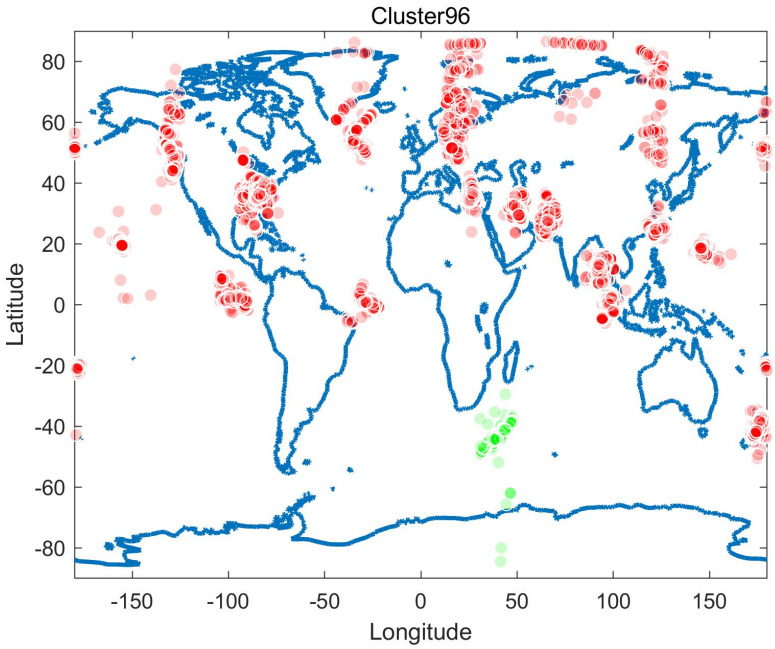


Figure 96: Global earthquakes causal relationship for target cluster96, highlighted in green, all other driving areas highlighted in red

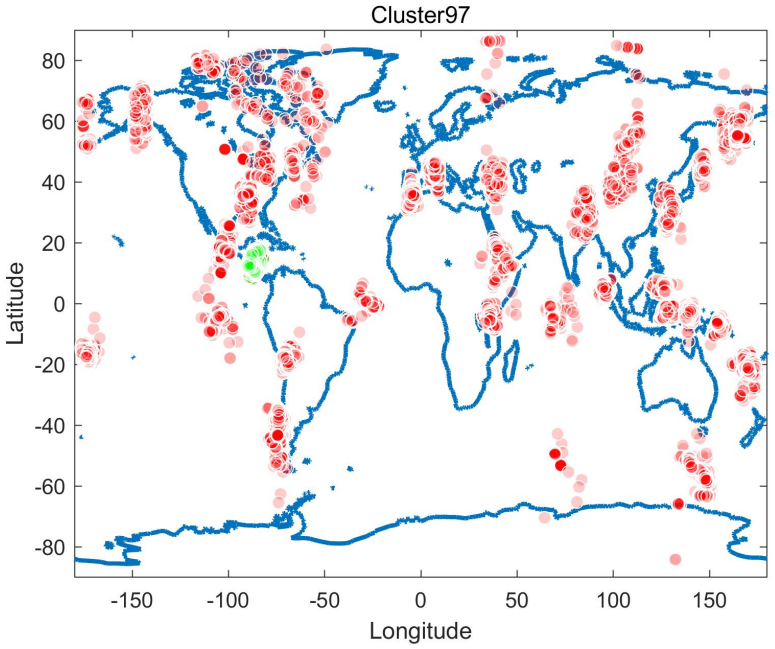


Figure 97: Global earthquakes causal relationship for target cluster97, highlighted in green, all other driving areas highlighted in red

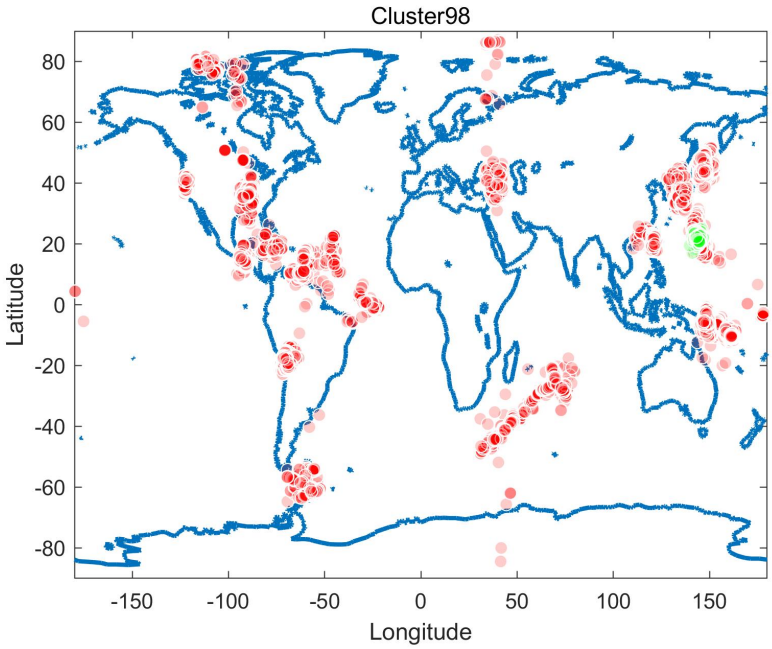


Figure 98: Global earthquakes causal relationship for target cluster98, highlighted in green, all other driving areas highlighted in red

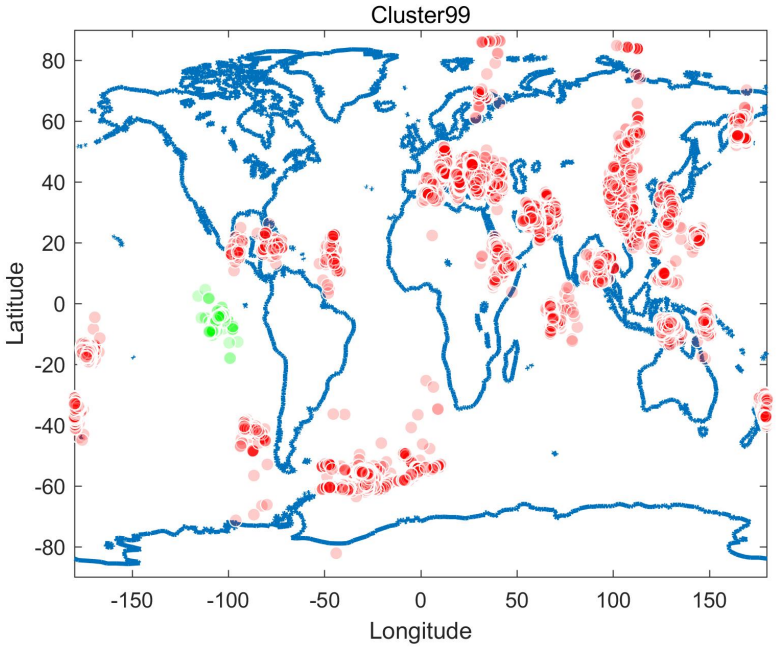


Figure 99: Global earthquakes causal relationship for target cluster99, highlighted in green, all other driving areas highlighted in red

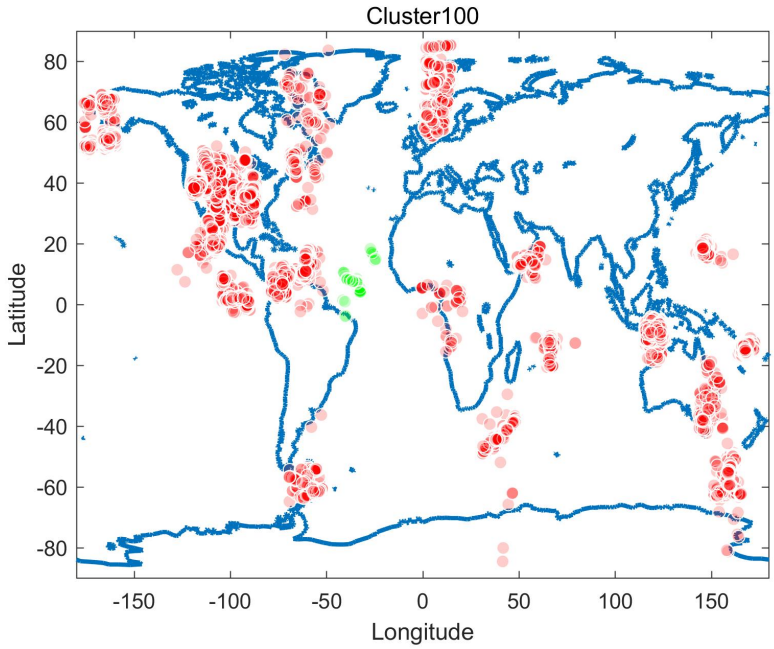


Figure 100: Global earthquakes causal relationship for target cluster100, highlighted in green, all other driving areas highlighted in red

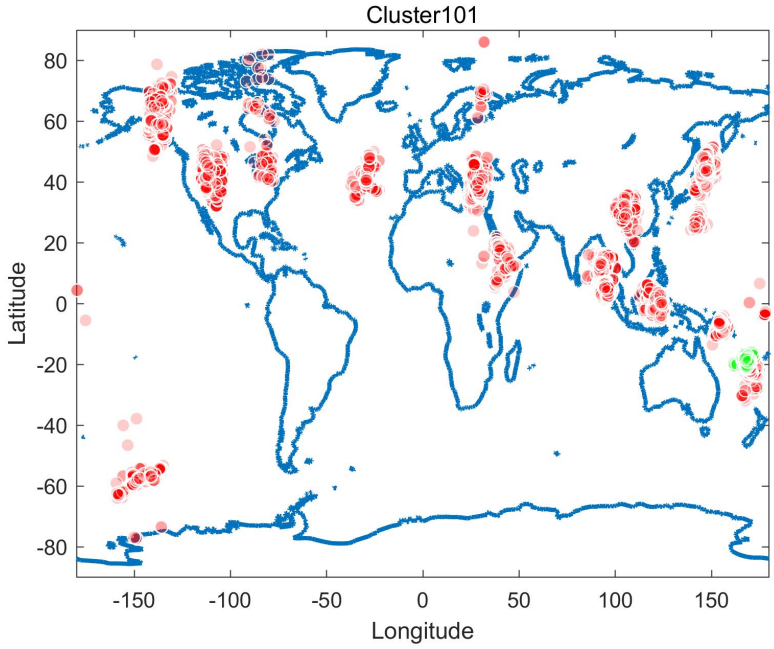


Figure 101: Global earthquakes causal relationship for target cluster101, highlighted in green, all other driving areas highlighted in red

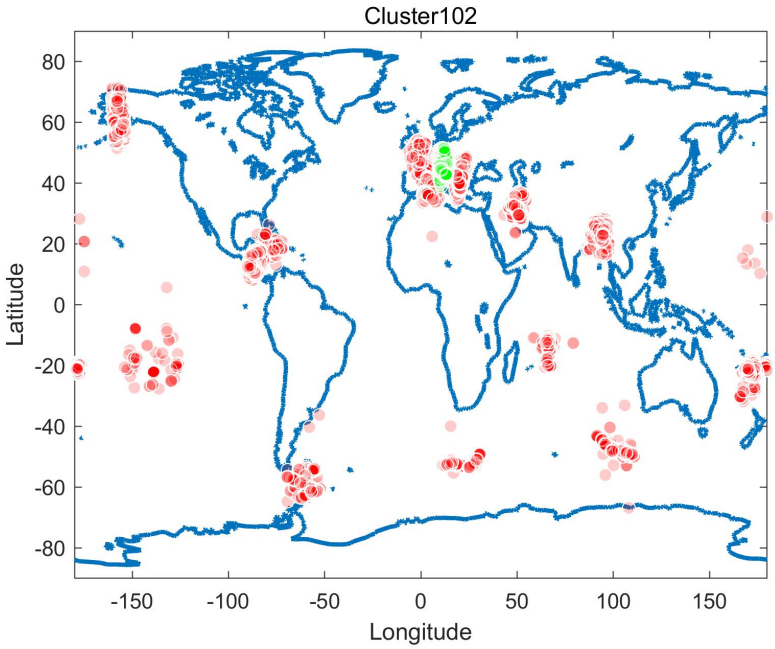


Figure 102: Global earthquakes causal relationship for target cluster102, highlighted in green, all other driving areas highlighted in red

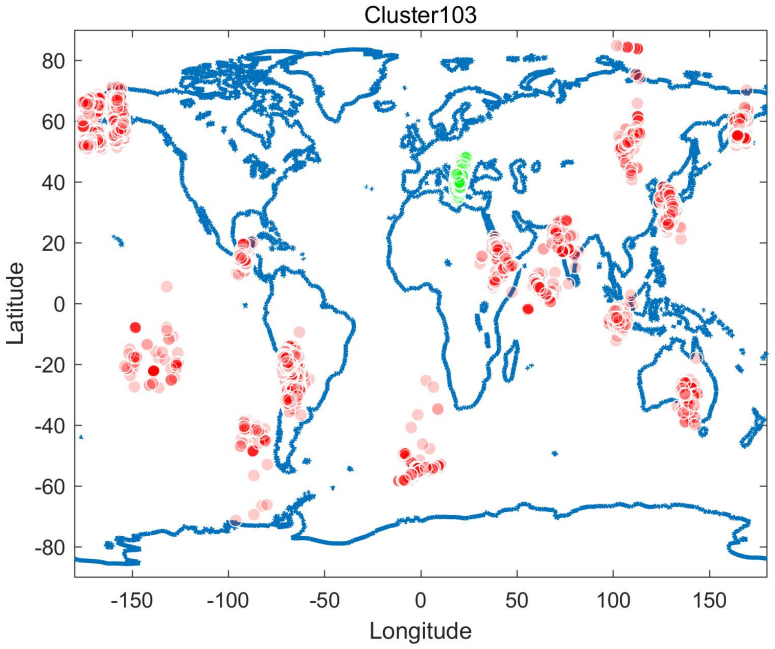


Figure 103: Global earthquakes causal relationship for target cluster103, highlighted in green, all other driving areas highlighted in red

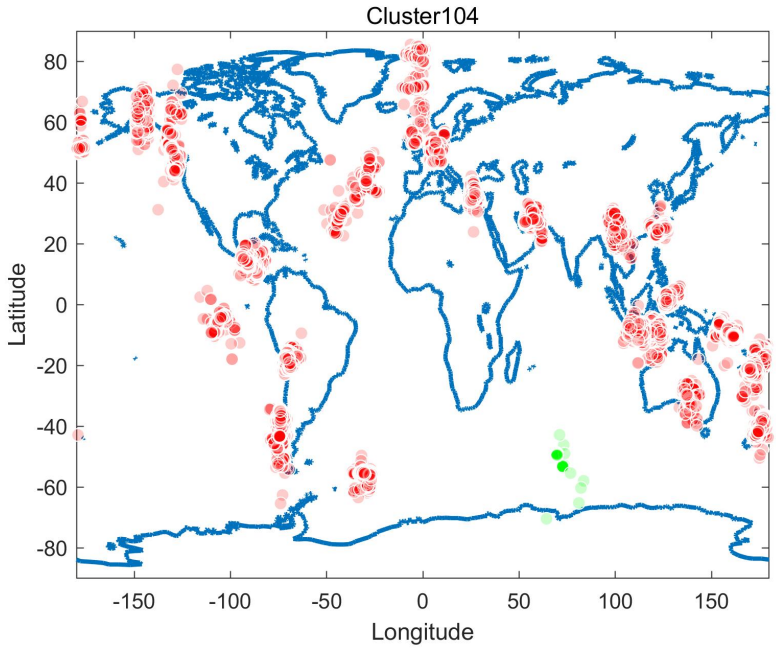


Figure 104: Global earthquakes causal relationship for target cluster104, highlighted in green, all other driving areas highlighted in red

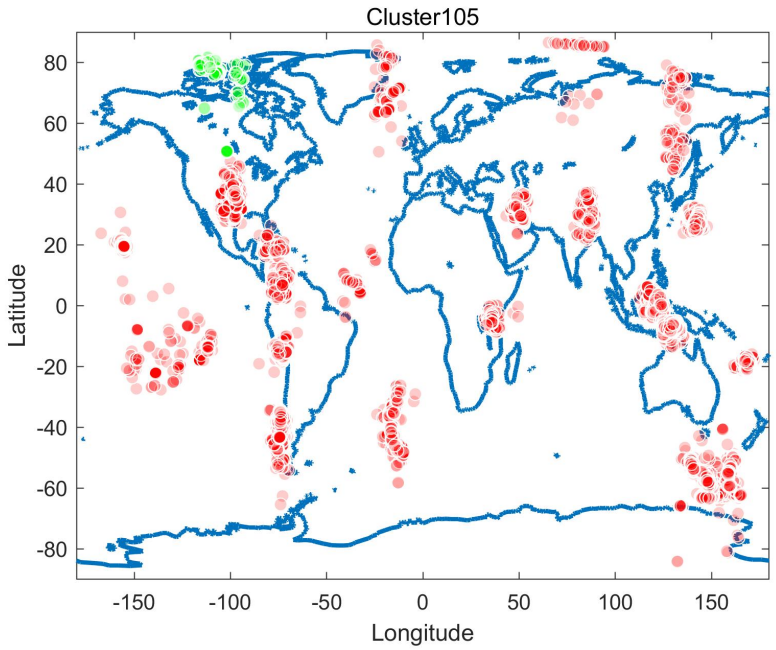


Figure 105: Global earthquakes causal relationship for target cluster105, highlighted in green, all other driving areas highlighted in red

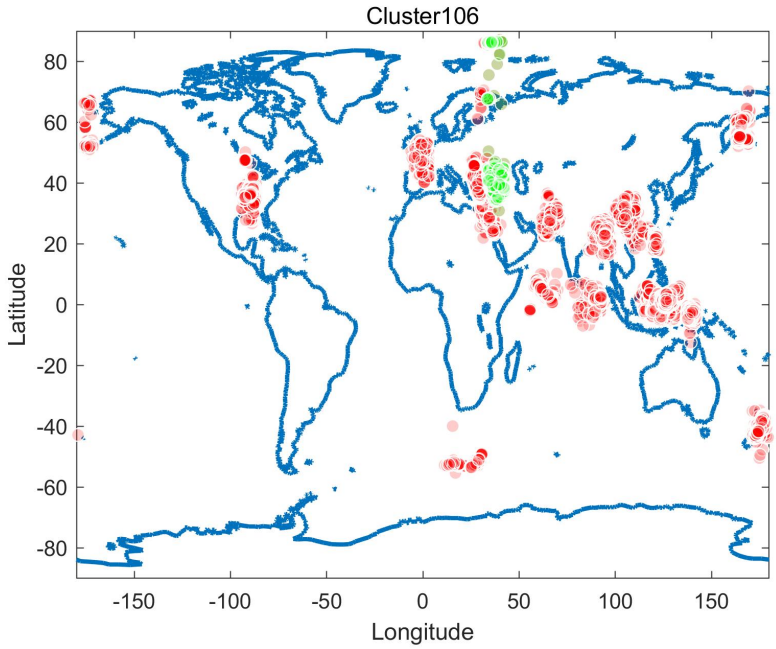


Figure 106: Global earthquakes causal relationship for target cluster106, highlighted in green, all other driving areas highlighted in red

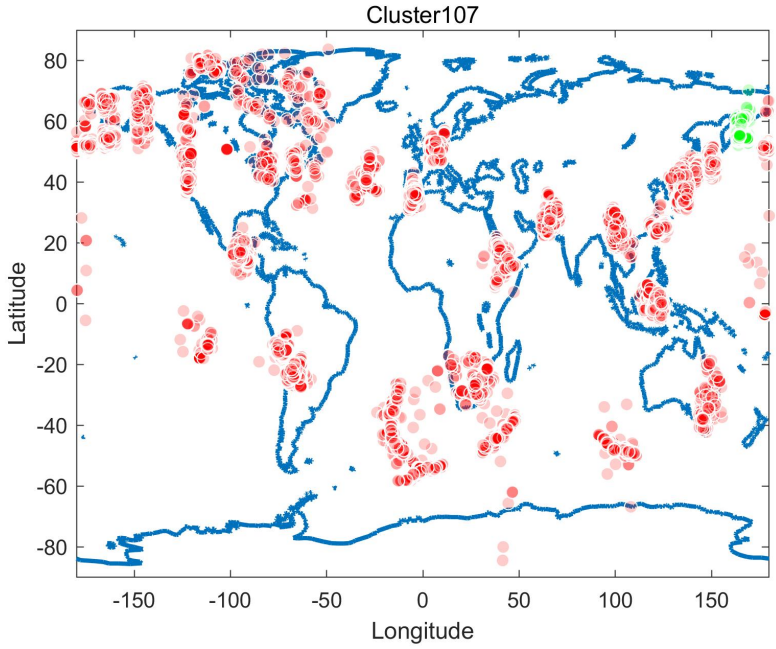


Figure 107: Global earthquakes causal relationship for target cluster107, highlighted in green, all other driving areas highlighted in red

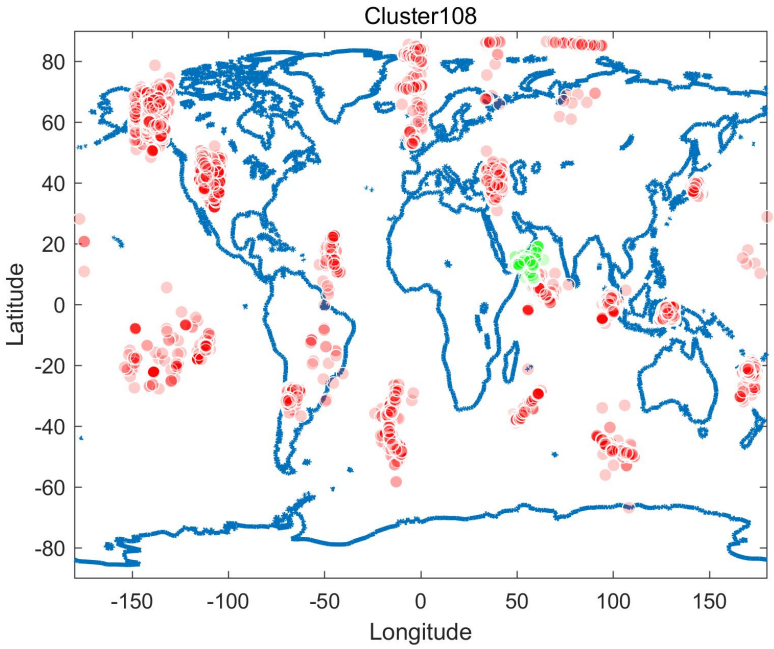


Figure 108: Global earthquakes causal relationship for target cluster108, highlighted in green, all other driving areas highlighted in red

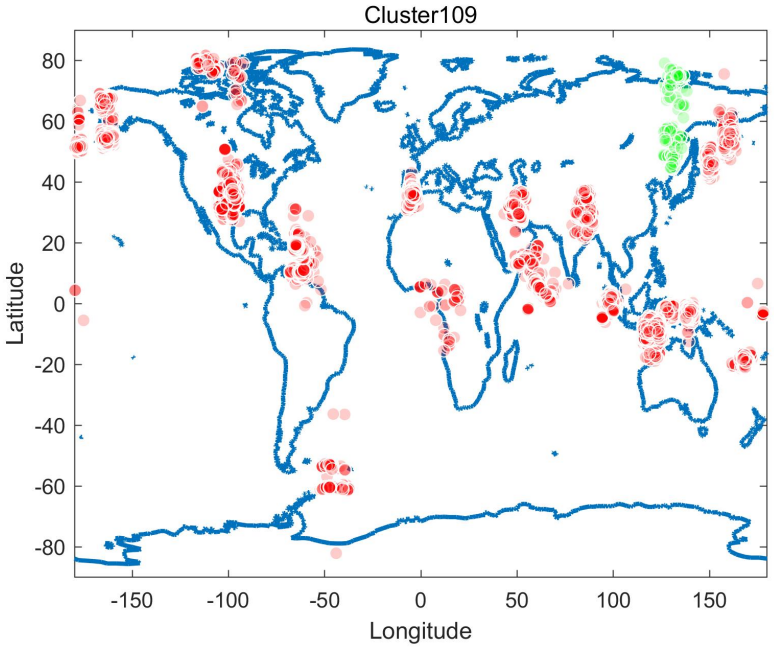


Figure 109: Global earthquakes causal relationship for target cluster109, highlighted in green, all other driving areas highlighted in red

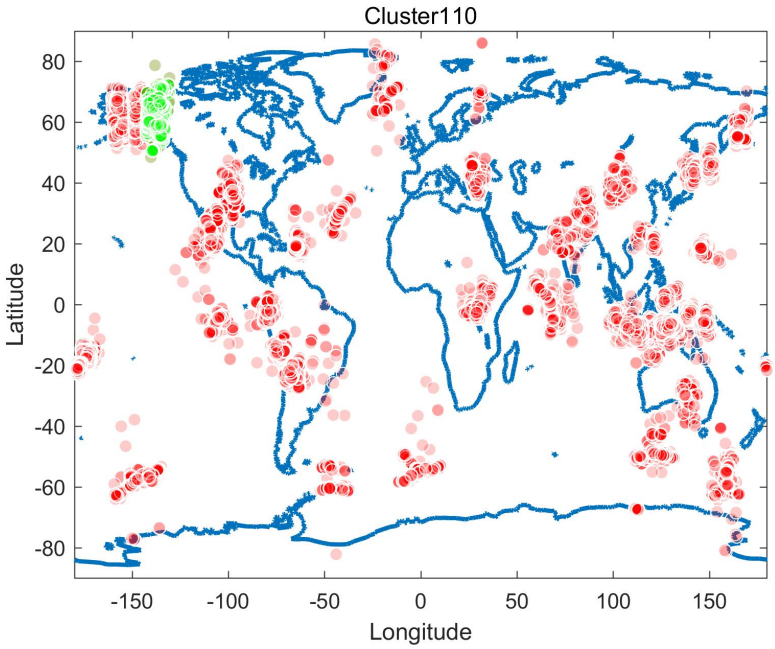


Figure 110: Global earthquakes causal relationship for target cluster110, highlighted in green, all other driving areas highlighted in red

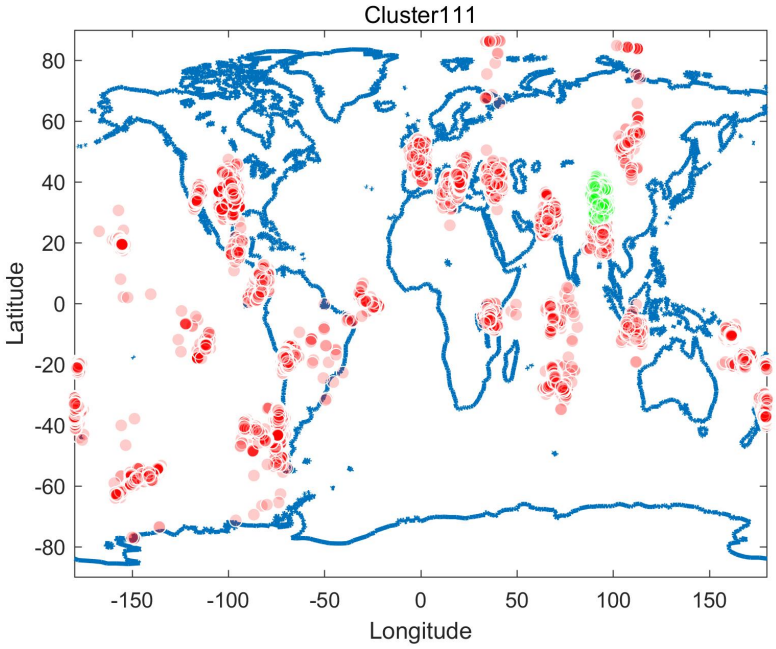


Figure 111: Global earthquakes causal relationship for target cluster111, highlighted in green, all other driving areas highlighted in red

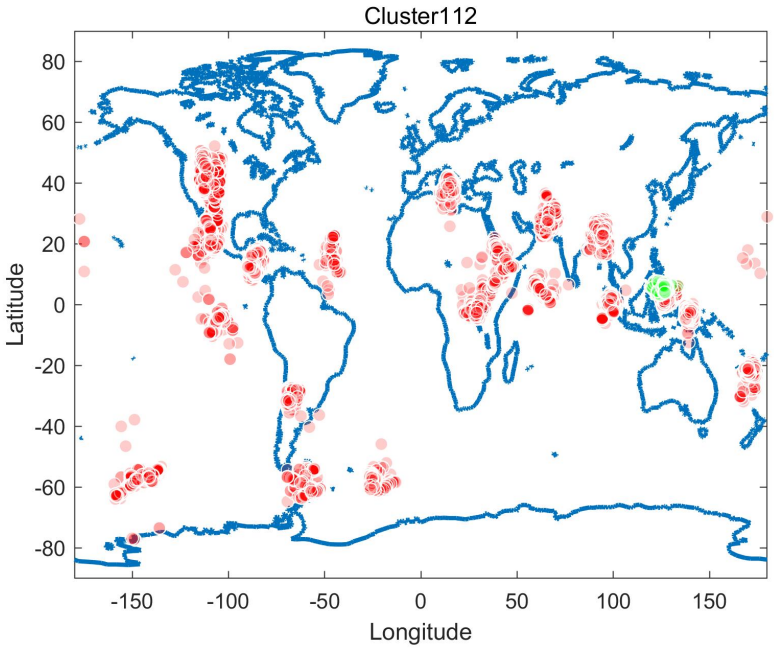


Figure 112: Global earthquakes causal relationship for target cluster112, highlighted in green, all other driving areas highlighted in red

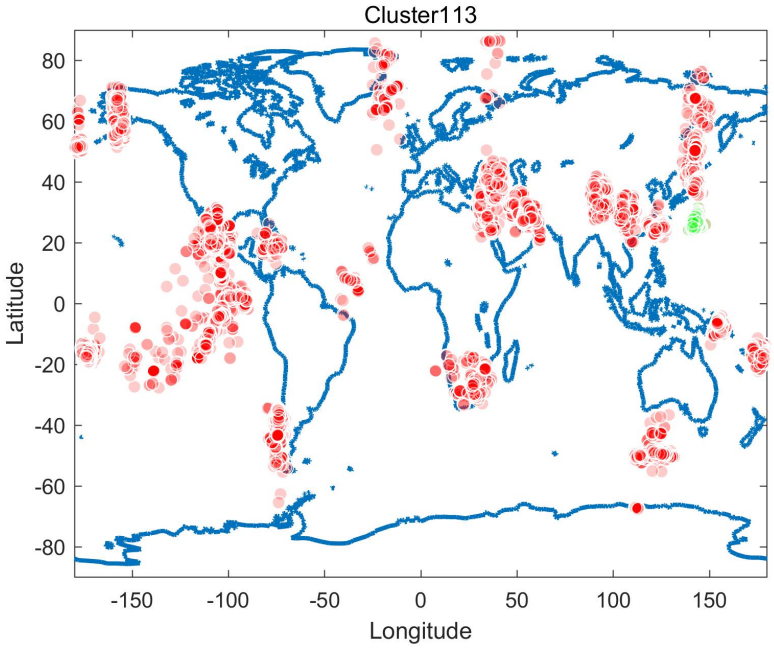


Figure 113: Global earthquakes causal relationship for target cluster113, highlighted in green, all other driving areas highlighted in red

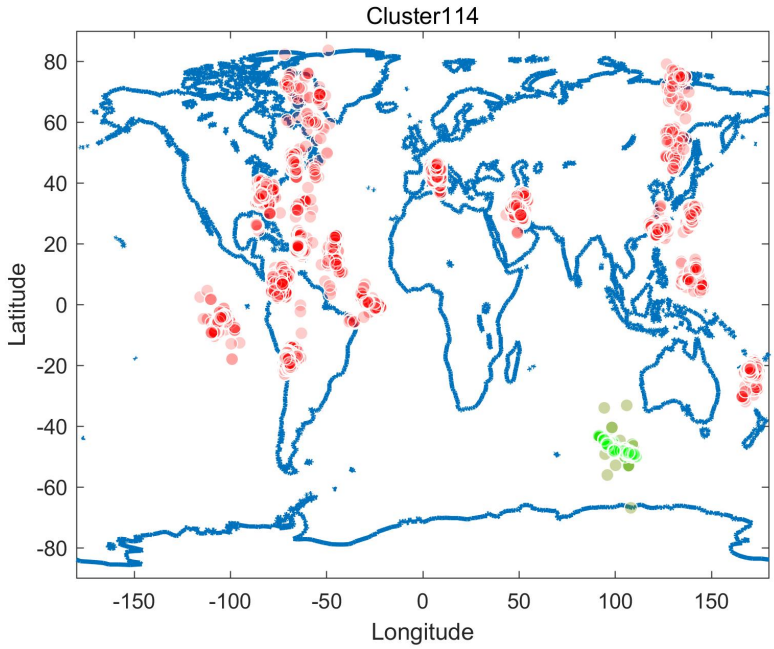


Figure 114: Global earthquakes causal relationship for target cluster114, highlighted in green, all other driving areas highlighted in red

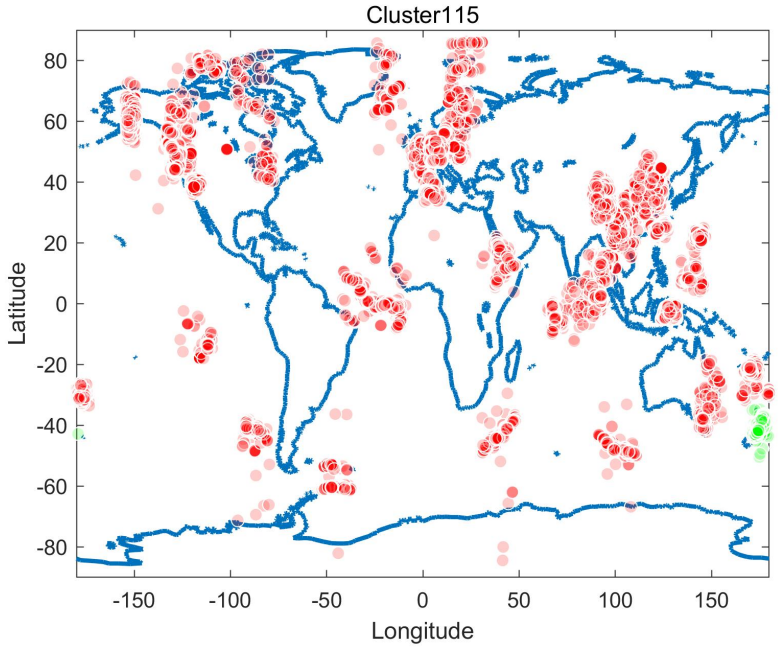


Figure 115: Global earthquakes causal relationship for target cluster115, highlighted in green, all other driving areas highlighted in red

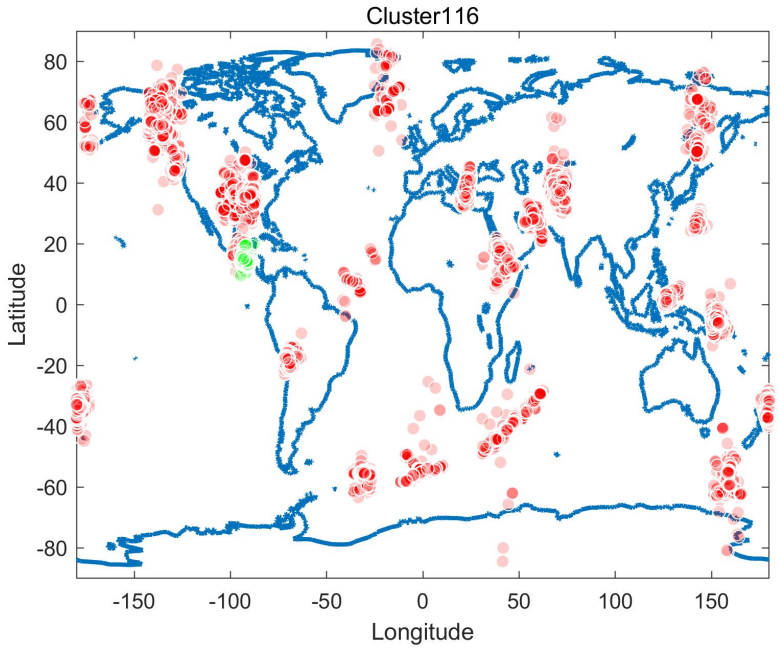


Figure 116: Global earthquakes causal relationship for target cluster116, highlighted in green, all other driving areas highlighted in red

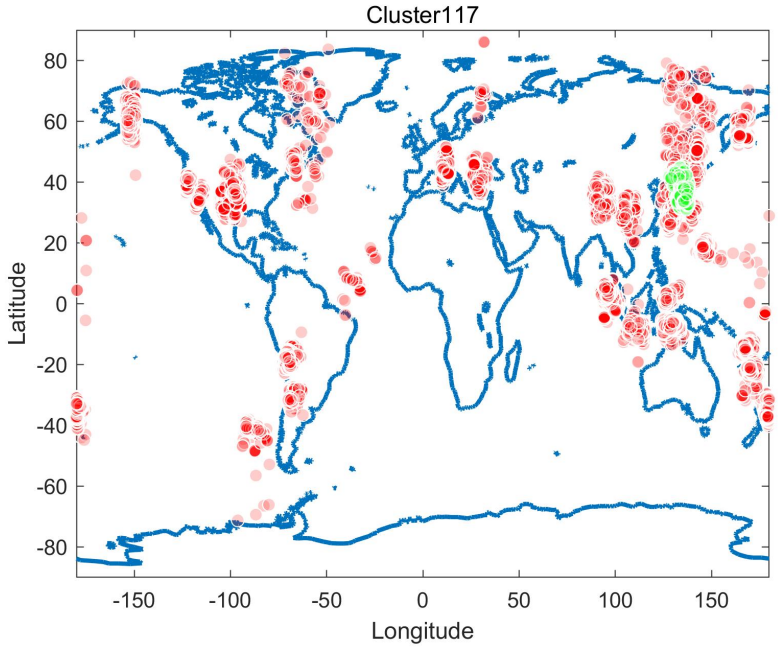


Figure 117: Global earthquakes causal relationship for target cluster117, highlighted in green, all other driving areas highlighted in red

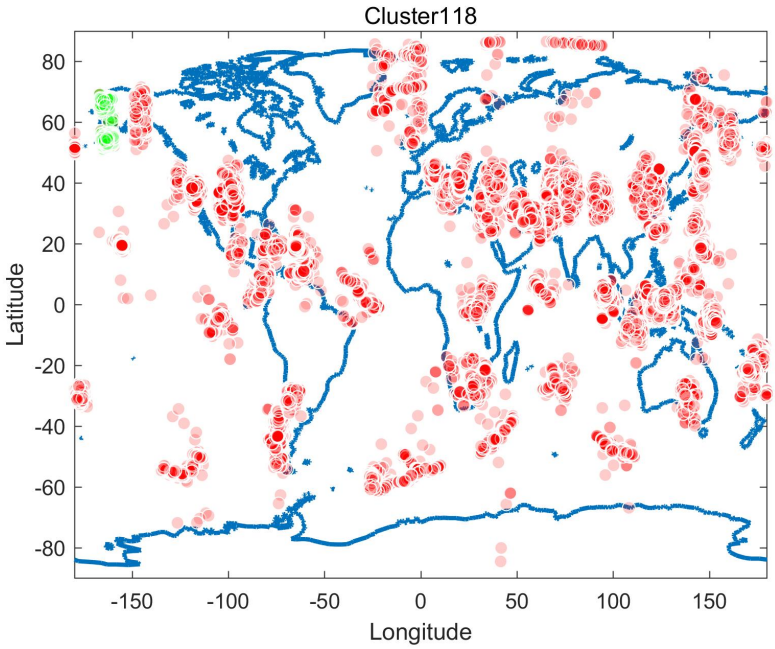


Figure 118: Global earthquakes causal relationship for target cluster118, highlighted in green, all other driving areas highlighted in red

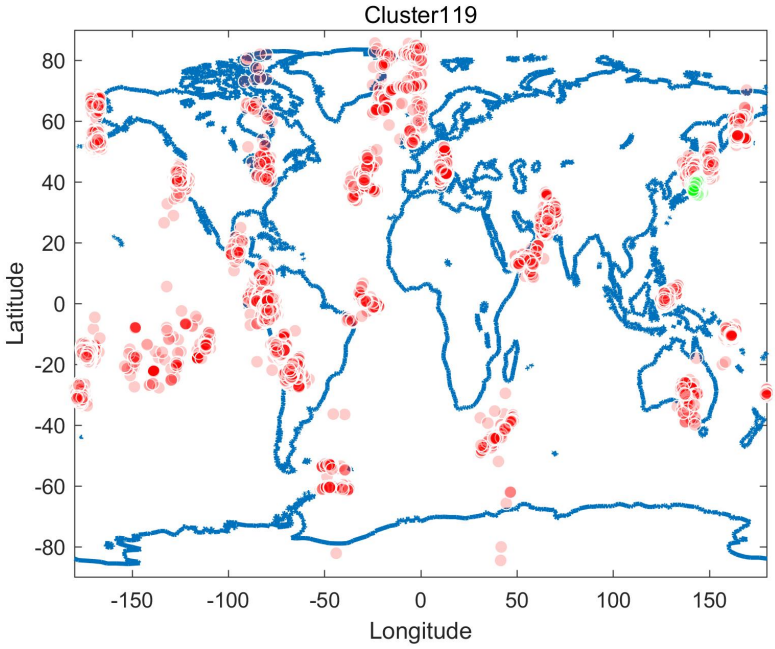


Figure 119: Global earthquakes causal relationship for target cluster119, highlighted in green, all other driving areas highlighted in red

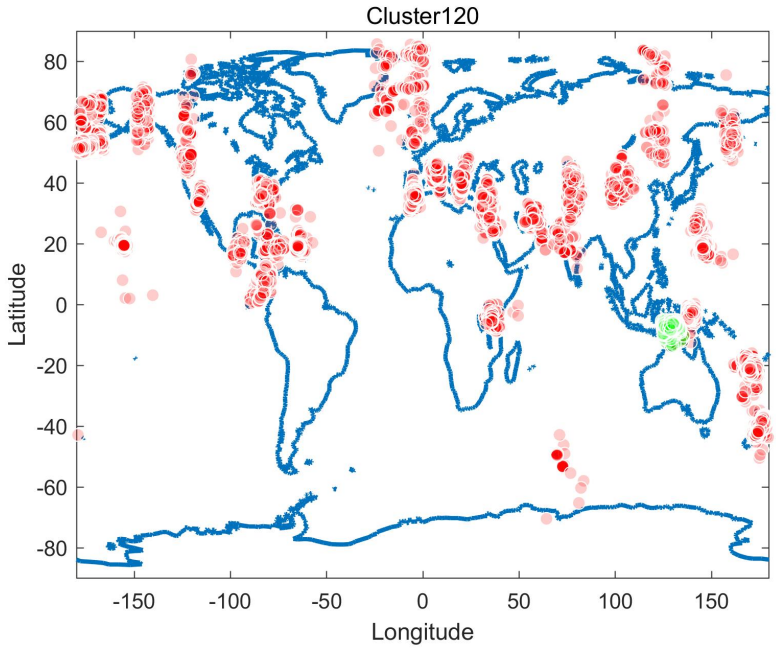


Figure 120: Global earthquakes causal relationship for target cluster120, highlighted in green, all other driving areas highlighted in red

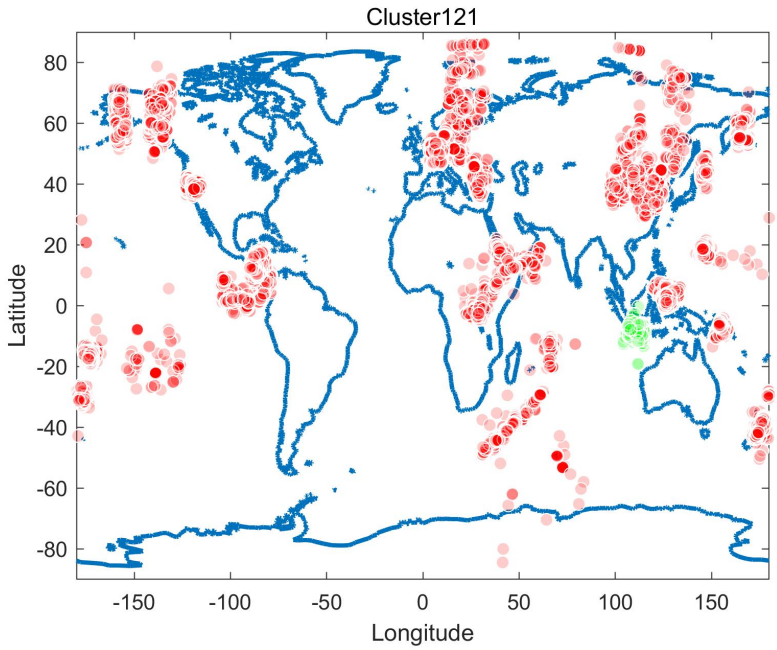


Figure 121: Global earthquakes causal relationship for target cluster121, highlighted in green, all other driving areas highlighted in red

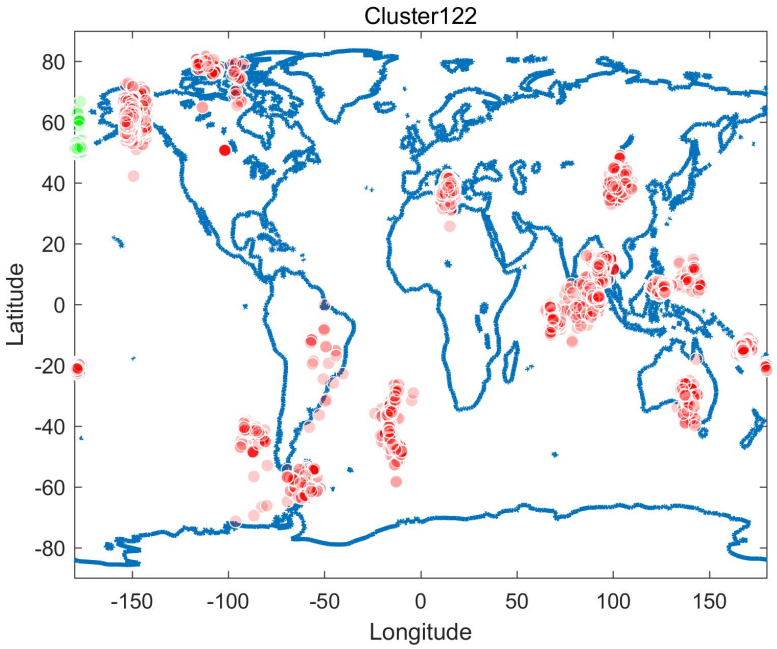


Figure 122: Global earthquakes causal relationship for target cluster122, highlighted in green, all other driving areas highlighted in red

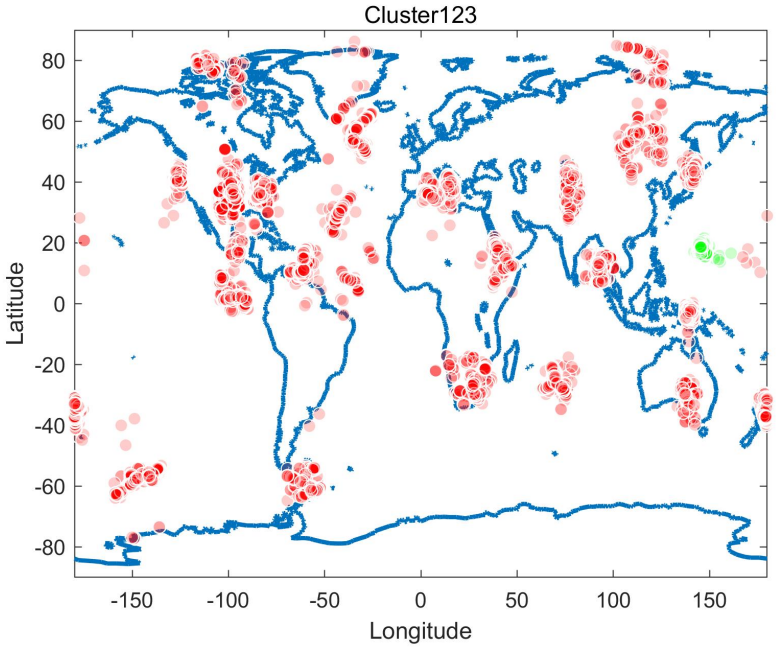


Figure 123: Global earthquakes causal relationship for target cluster123, highlighted in green, all other driving areas highlighted in red

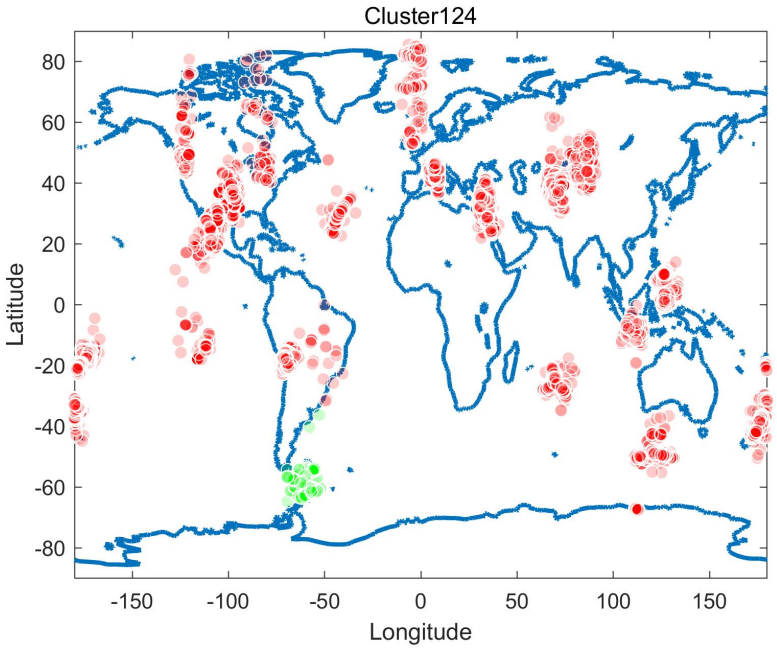


Figure 124: Global earthquakes causal relationship for target cluster124, highlighted in green, all other driving areas highlighted in red

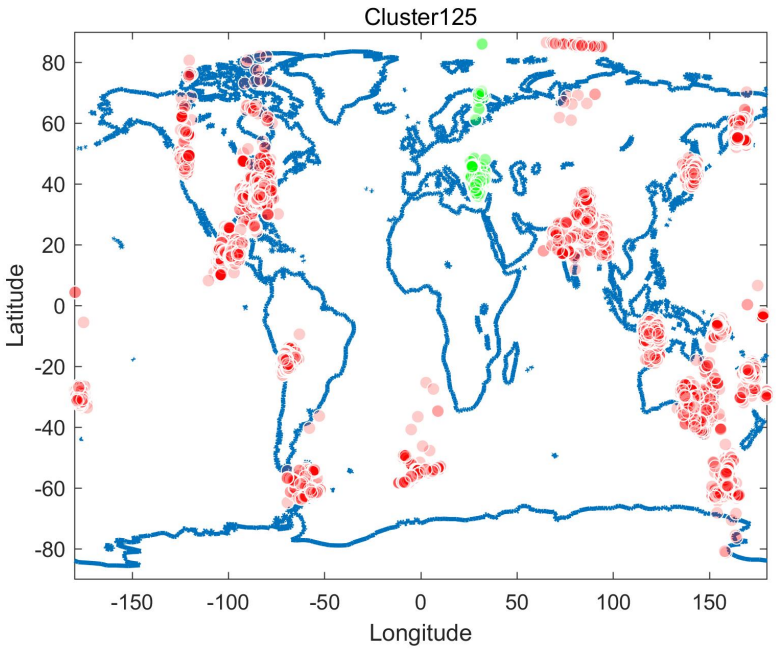


Figure 125: Global earthquakes causal relationship for target cluster125, highlighted in green, all other driving areas highlighted in red

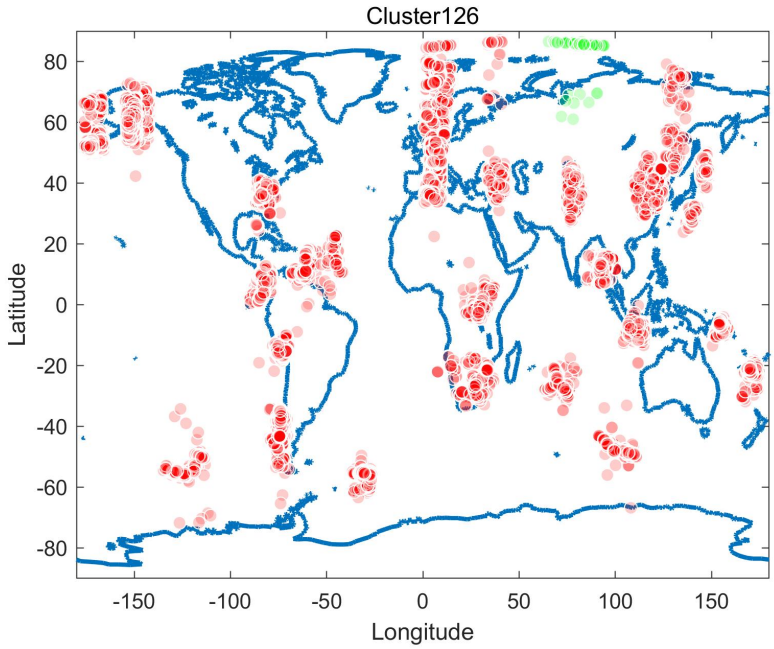


Figure 126: Global earthquakes causal relationship for target cluster126, highlighted in green, all other driving areas highlighted in red

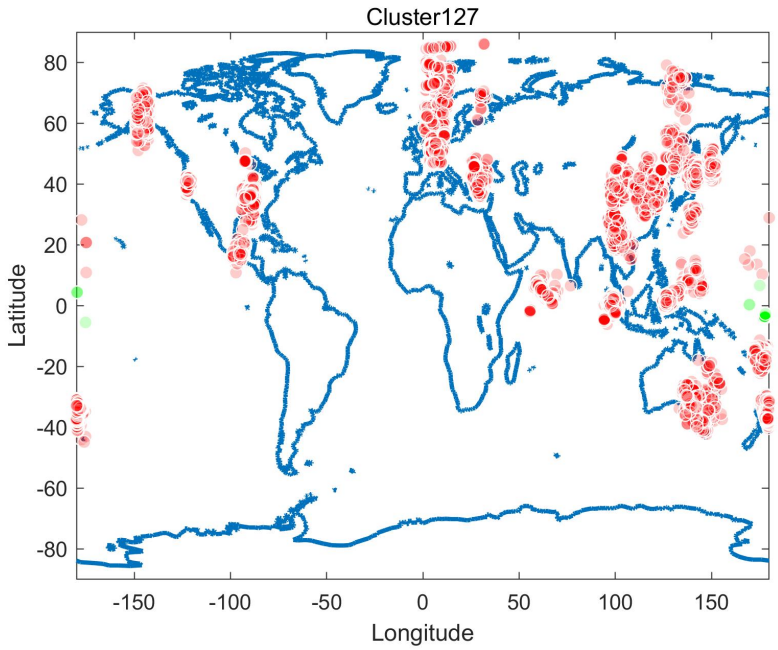


Figure 127: Global earthquakes causal relationship for target cluster127, highlighted in green, all other driving areas highlighted in red

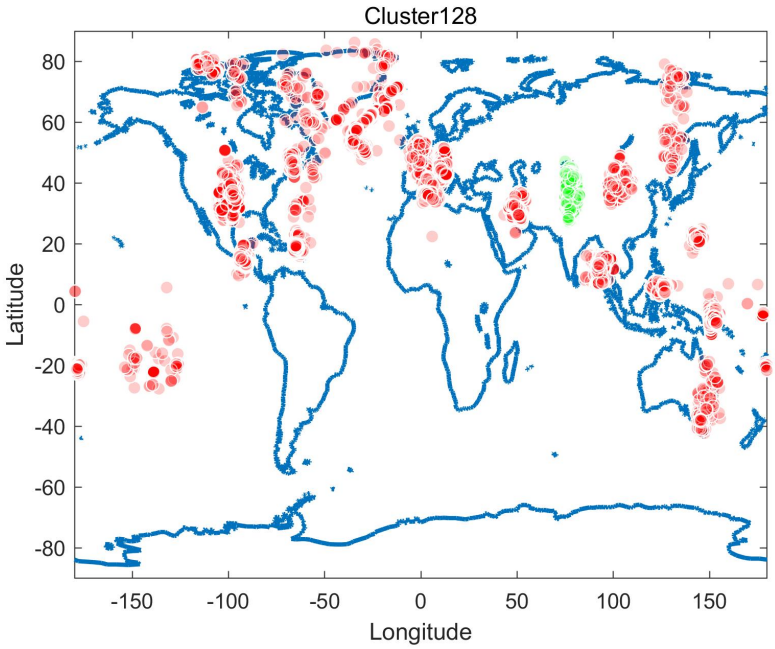


Figure 128: Global earthquakes causal relationship for target cluster128, highlighted in green, all other driving areas highlighted in red

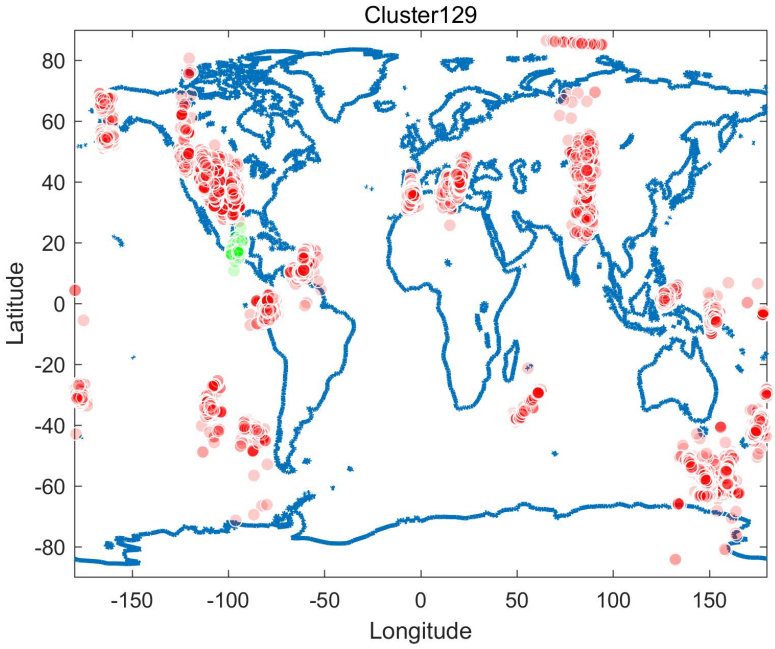


Figure 129: Global earthquakes causal relationship for target cluster129, highlighted in green, all other driving areas highlighted in red

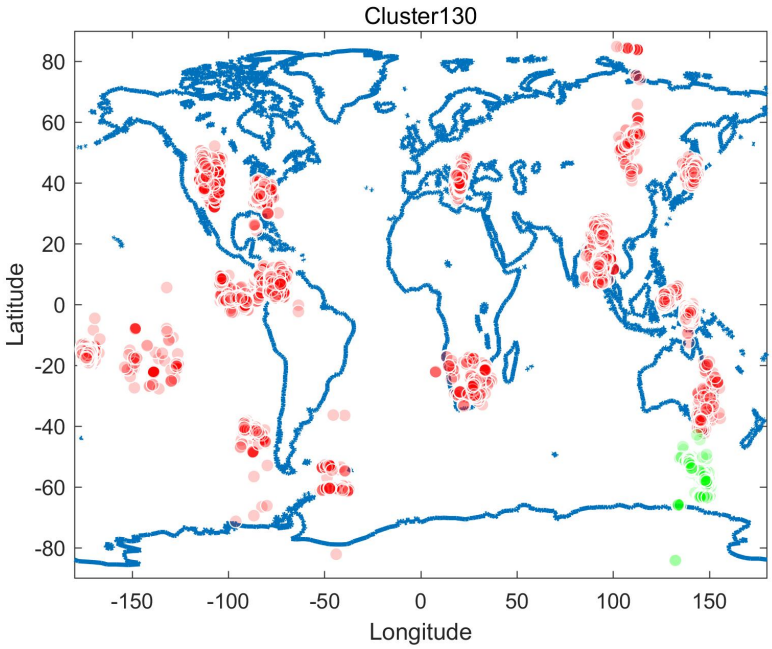


Figure 130: Global earthquakes causal relationship for target cluster130, highlighted in green, all other driving areas highlighted in red

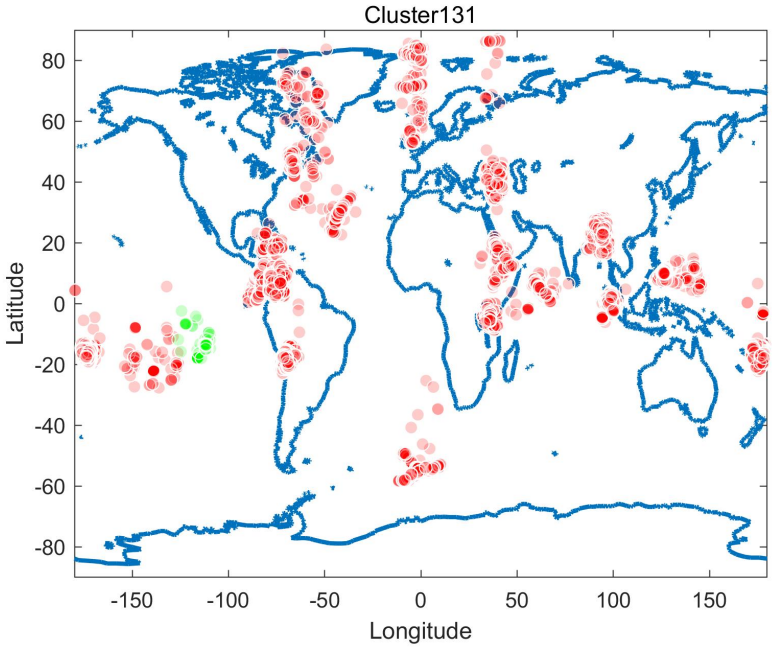


Figure 131: Global earthquakes causal relationship for target cluster131, highlighted in green, all other driving areas highlighted in red

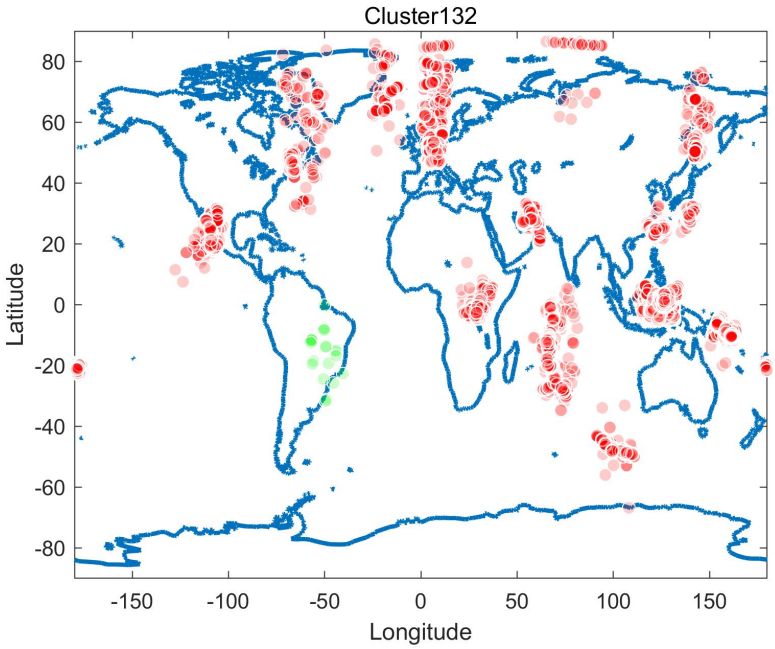


Figure 132: Global earthquakes causal relationship for target cluster132, highlighted in green, all other driving areas highlighted in red

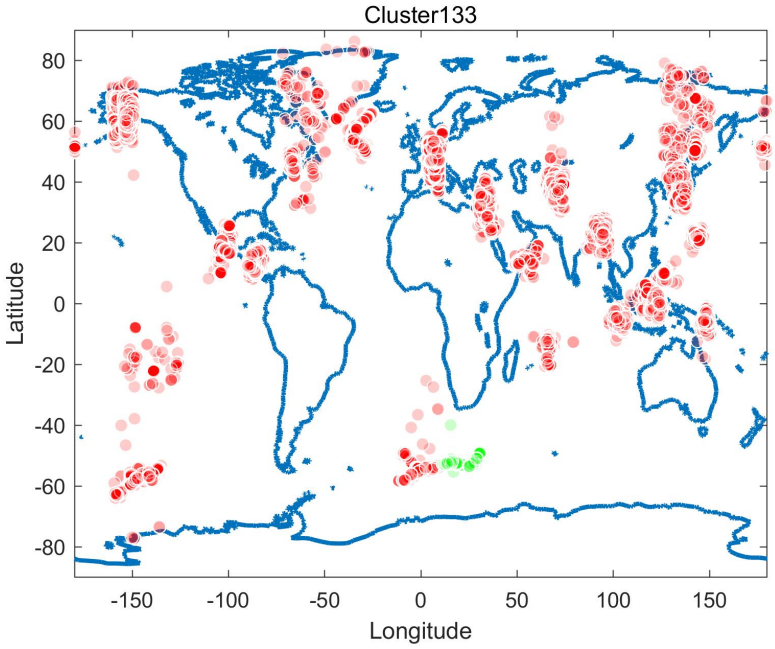


Figure 133: Global earthquakes causal relationship for target cluster133, highlighted in green, all other driving areas highlighted in red

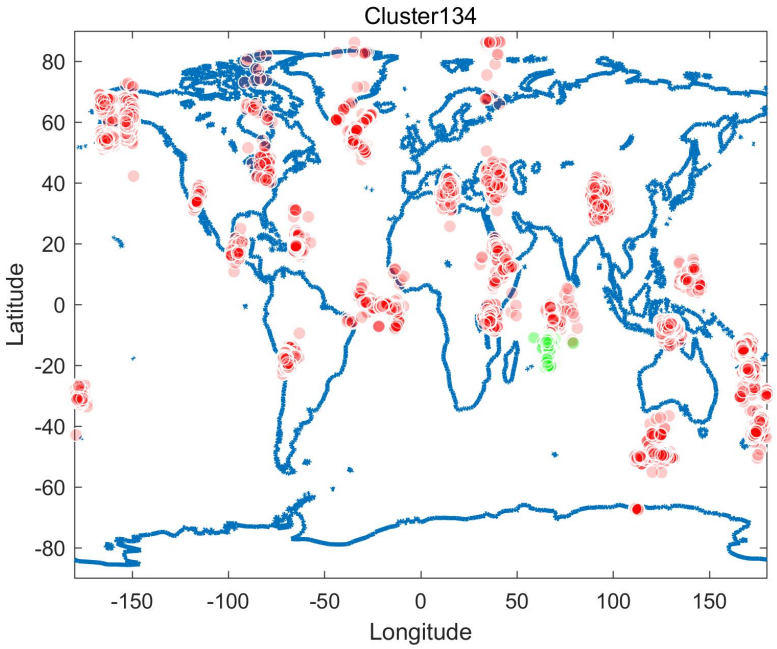


Figure 134: Global earthquakes causal relationship for target cluster134, highlighted in green, all other driving areas highlighted in red

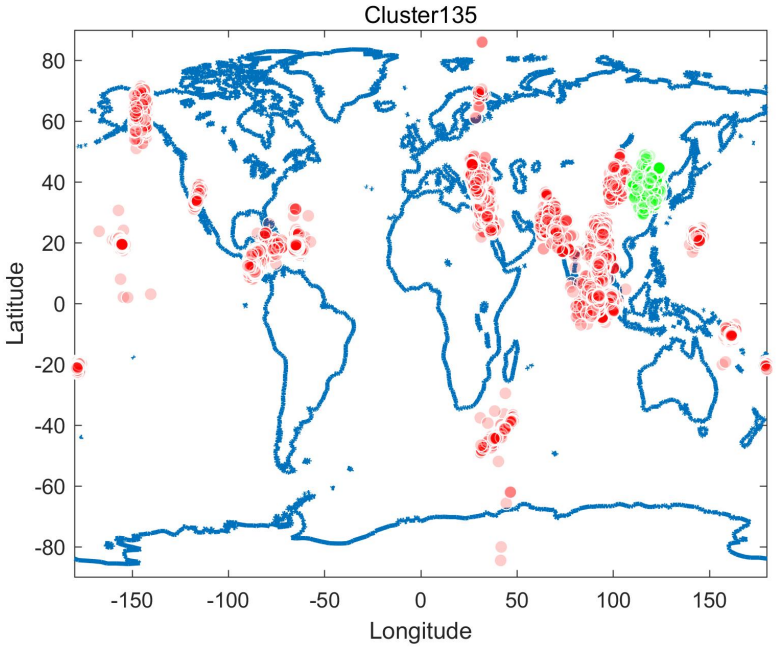


Figure 135: Global earthquakes causal relationship for target cluster135, highlighted in green, all other driving areas highlighted in red

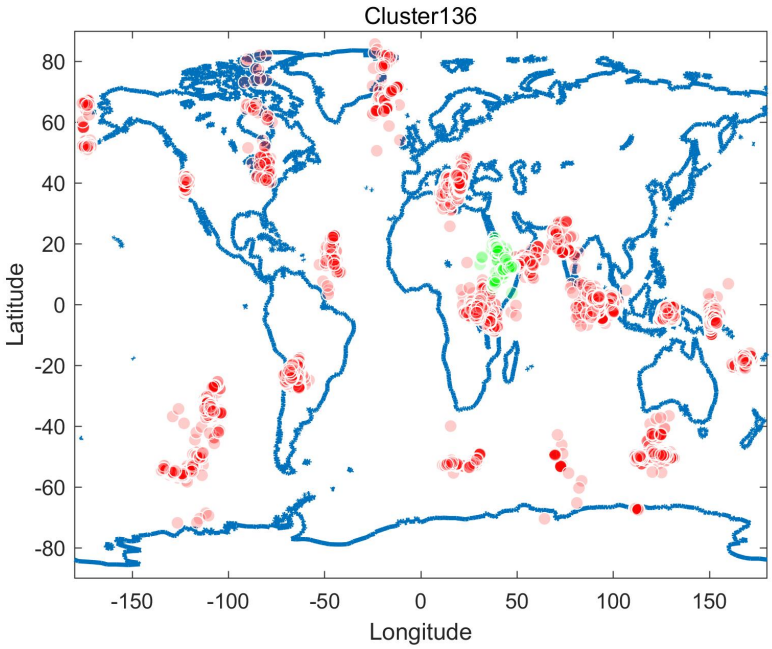


Figure 136: Global earthquakes causal relationship for target cluster136, highlighted in green, all other driving areas highlighted in red

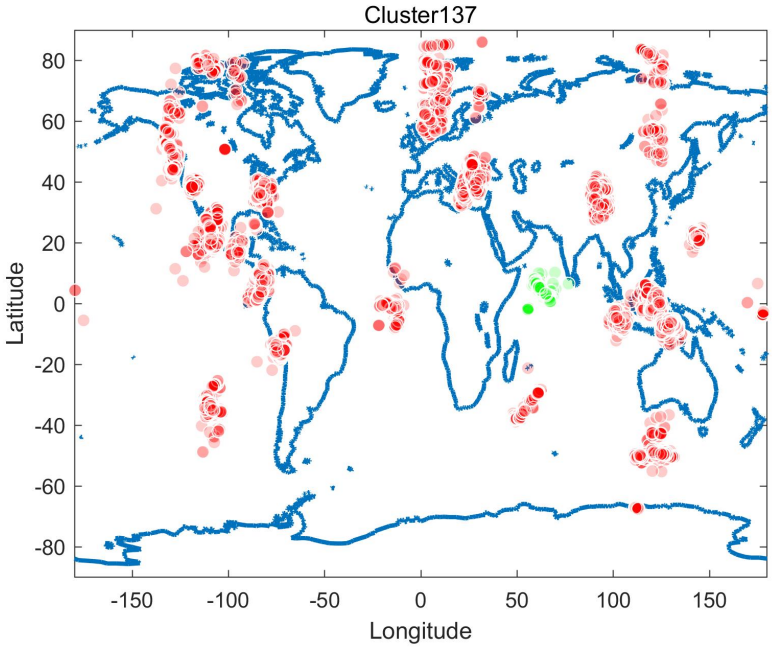


Figure 137: Global earthquakes causal relationship for target cluster137, highlighted in green, all other driving areas highlighted in red

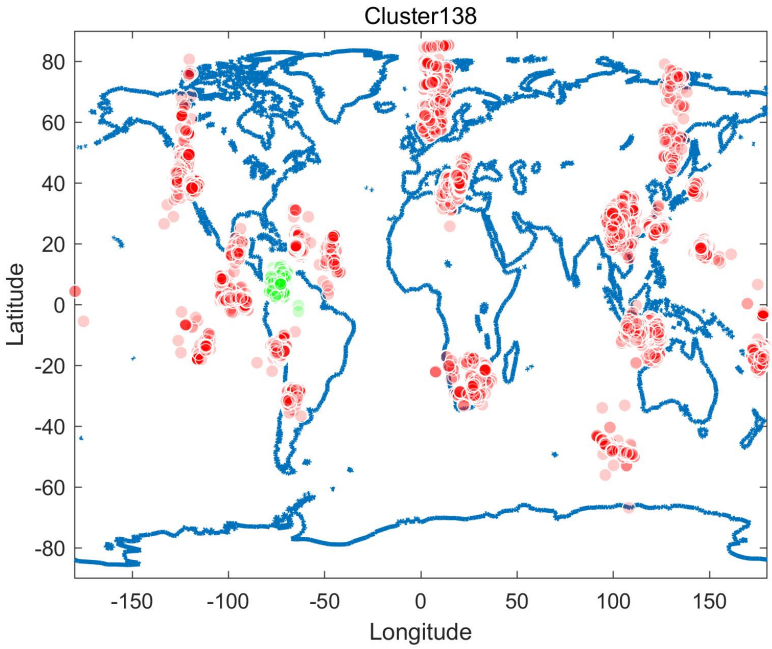


Figure 138: Global earthquakes causal relationship for target cluster138, highlighted in green, all other driving areas highlighted in red

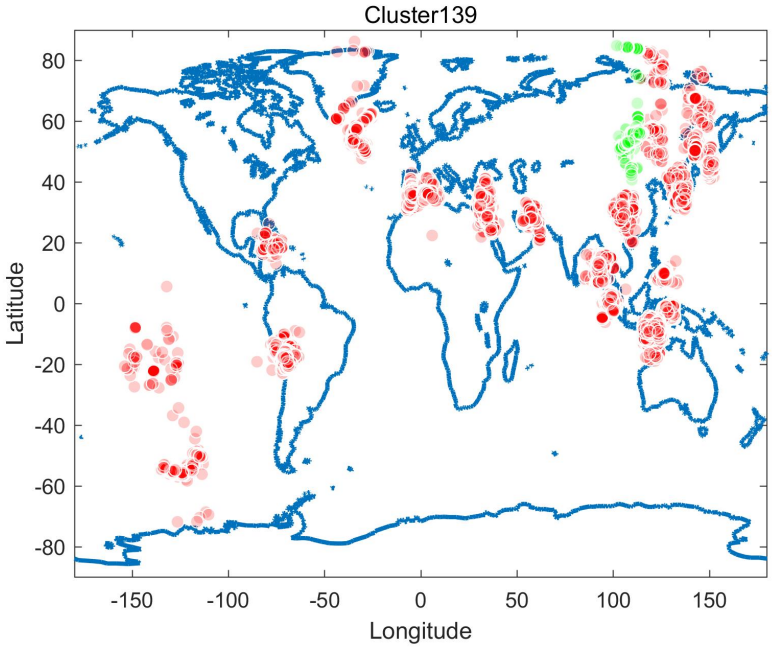


Figure 139: Global earthquakes causal relationship for target cluster139, highlighted in green, all other driving areas highlighted in red

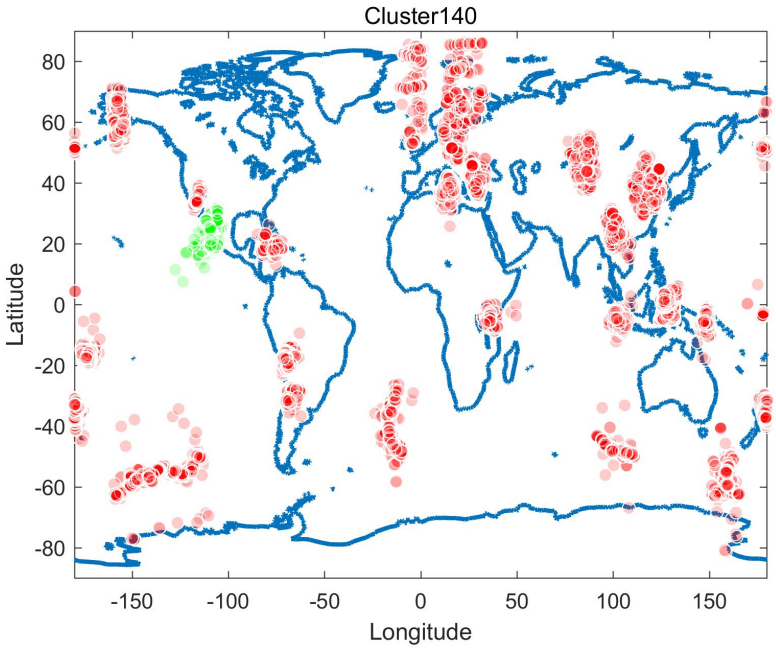


Figure 140: Global earthquakes causal relationship for target cluster140, highlighted in green, all other driving areas highlighted in red

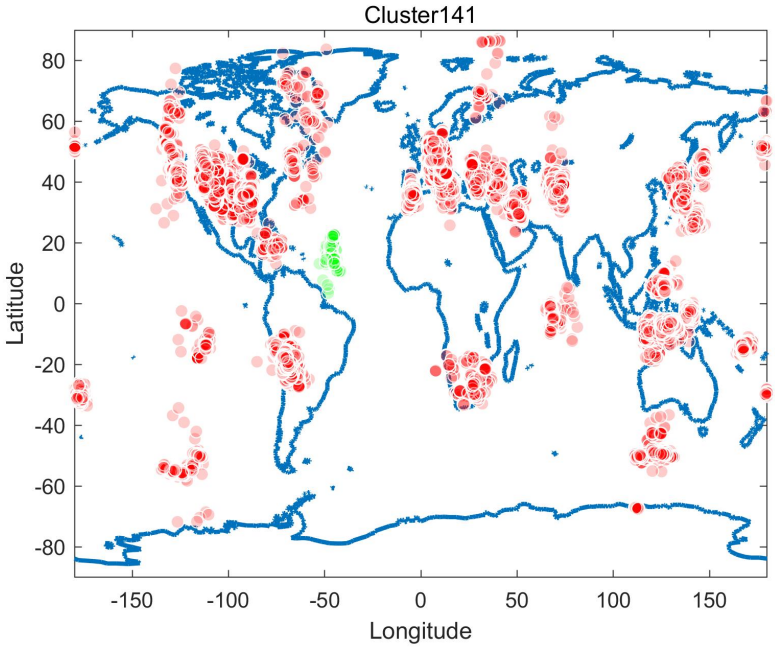


Figure 141: Global earthquakes causal relationship for target cluster141, highlighted in green, all other driving areas highlighted in red

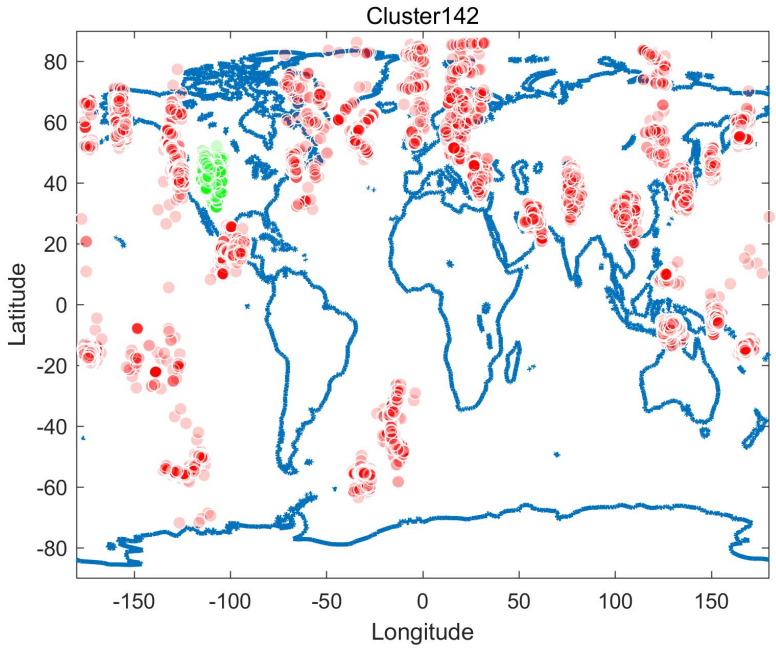


Figure 142: Global earthquakes causal relationship for target cluster142, highlighted in green, all other driving areas highlighted in red

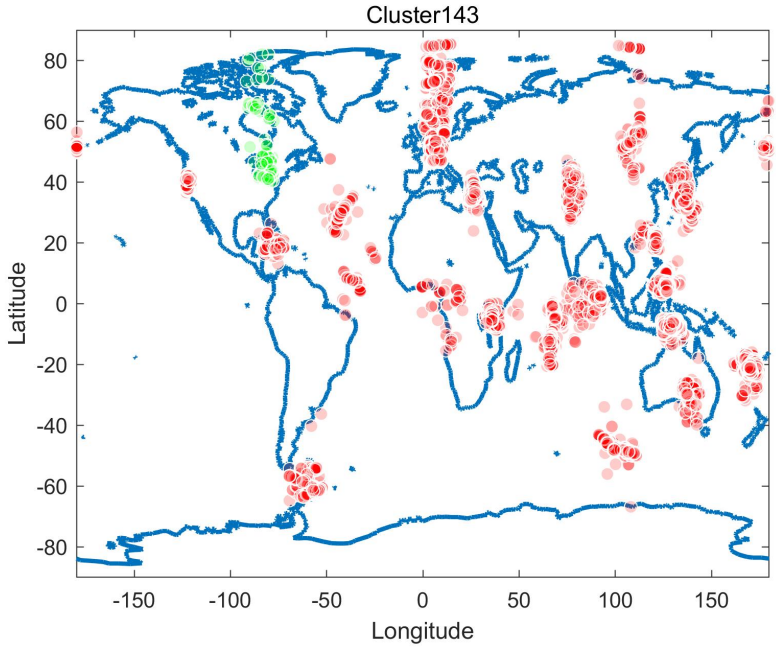


Figure 143: Global earthquakes causal relationship for target cluster143, highlighted in green, all other driving areas highlighted in red

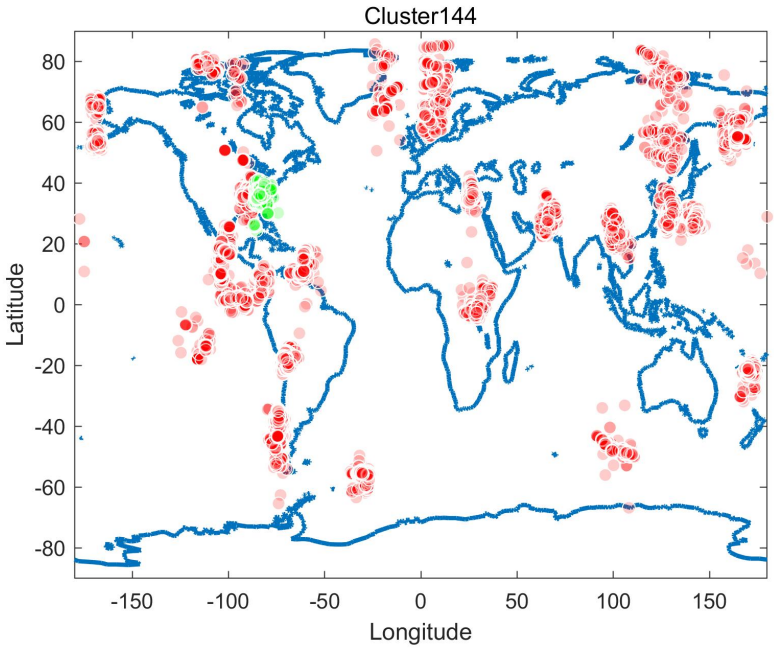


Figure 144: Global earthquakes causal relationship for target cluster144, highlighted in green, all other driving areas highlighted in red

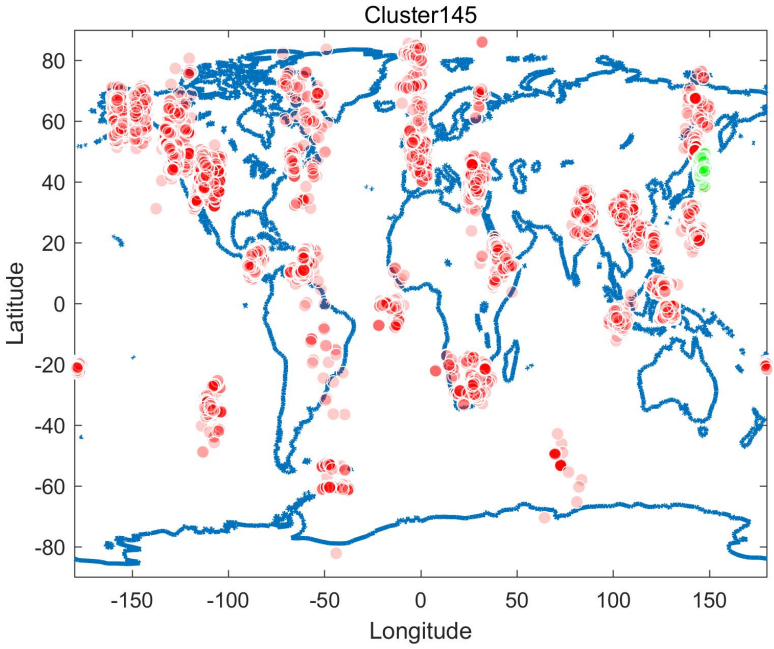


Figure 145: Global earthquakes causal relationship for target cluster145, highlighted in green, all other driving areas highlighted in red

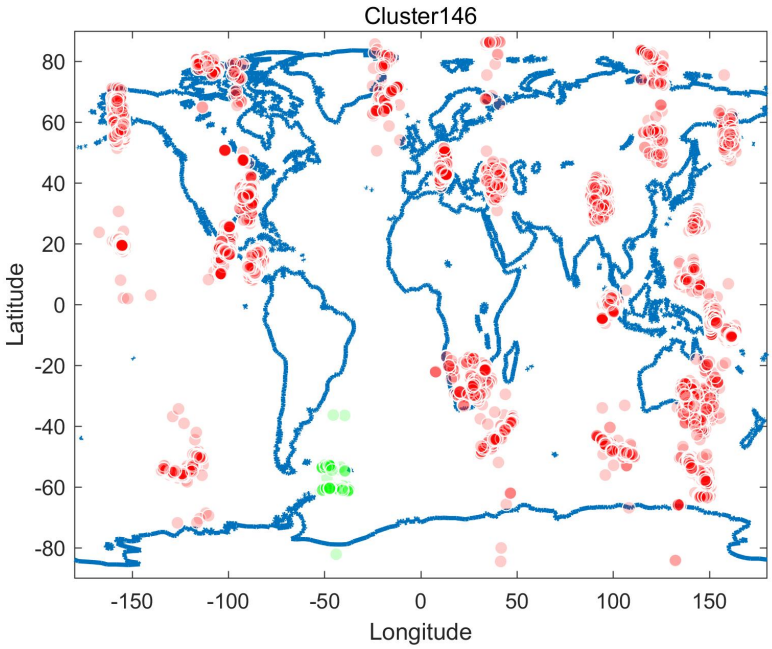


Figure 146: Global earthquakes causal relationship for target cluster146, highlighted in green, all other driving areas highlighted in red

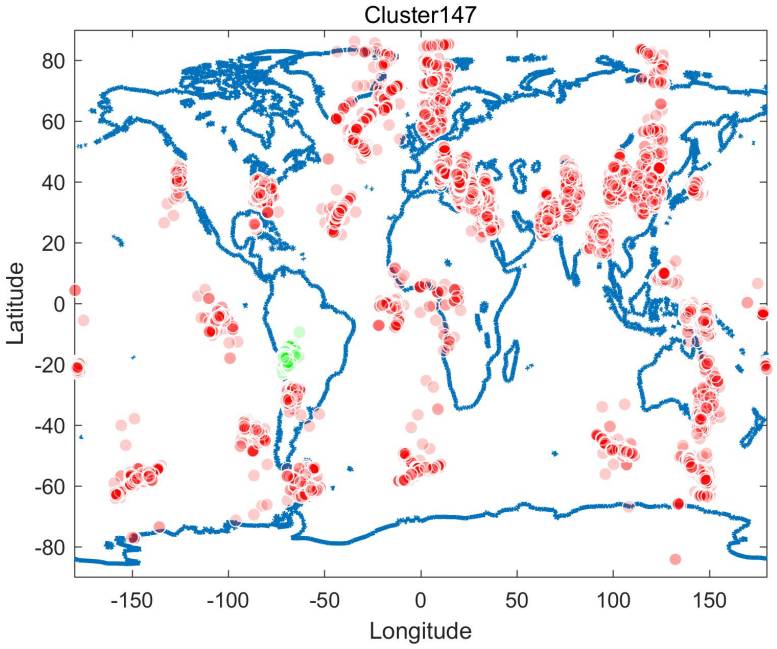


Figure 147: Global earthquakes causal relationship for target cluster147, highlighted in green, all other driving areas highlighted in red

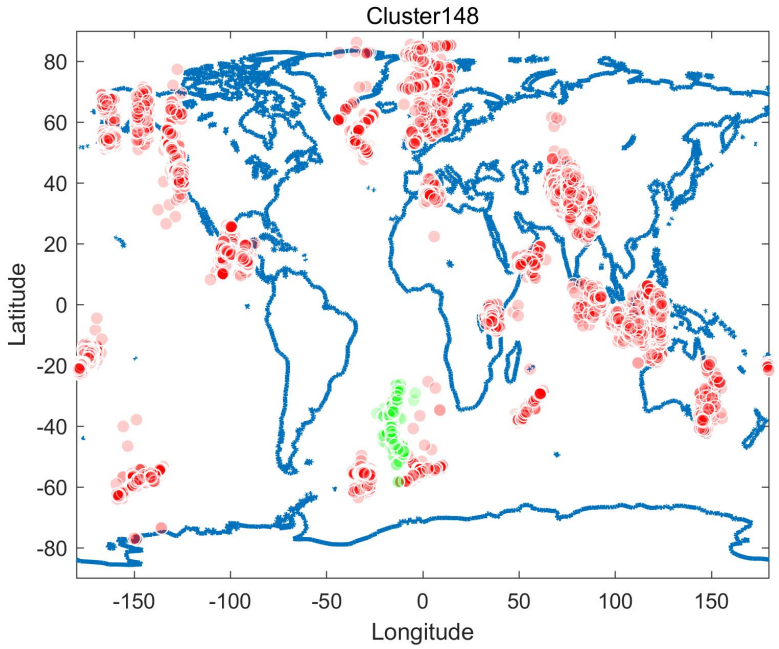


Figure 148: Global earthquakes causal relationship for target cluster148, highlighted in green, all other driving areas highlighted in red

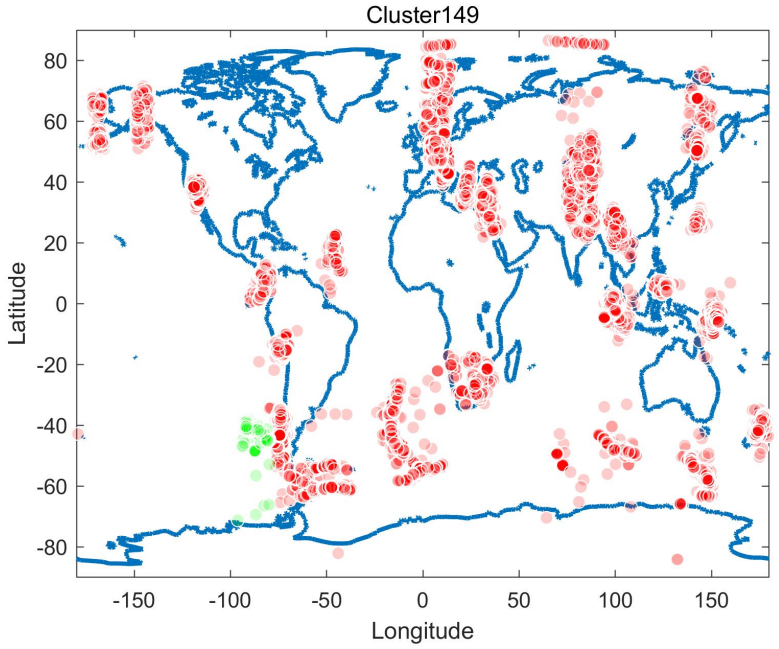


Figure 149: Global earthquakes causal relationship for target cluster149, highlighted in green, all other driving areas highlighted in red

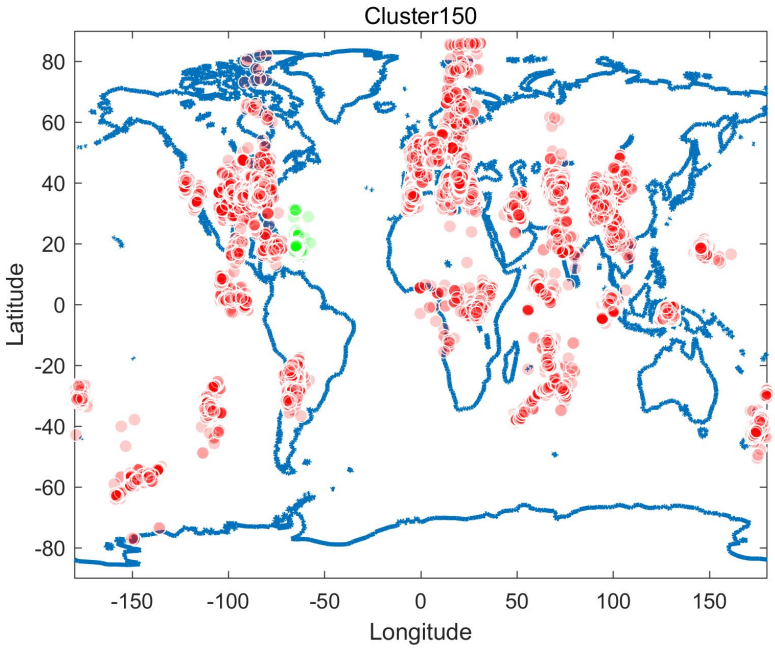


Figure 150: Global earthquakes causal relationship for target cluster150, highlighted in green, all other driving areas highlighted in red

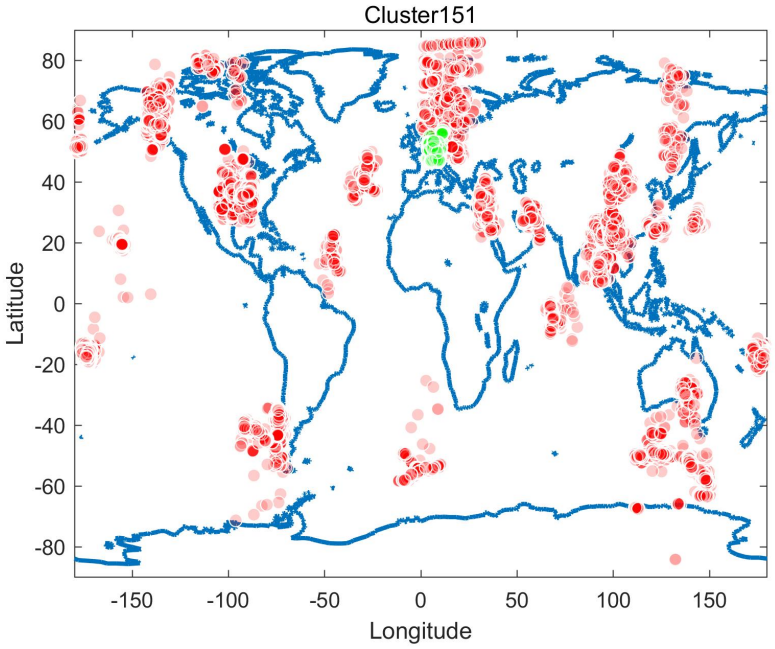


Figure 151: Global earthquakes causal relationship for target cluster151, highlighted in green, all other driving areas highlighted in red

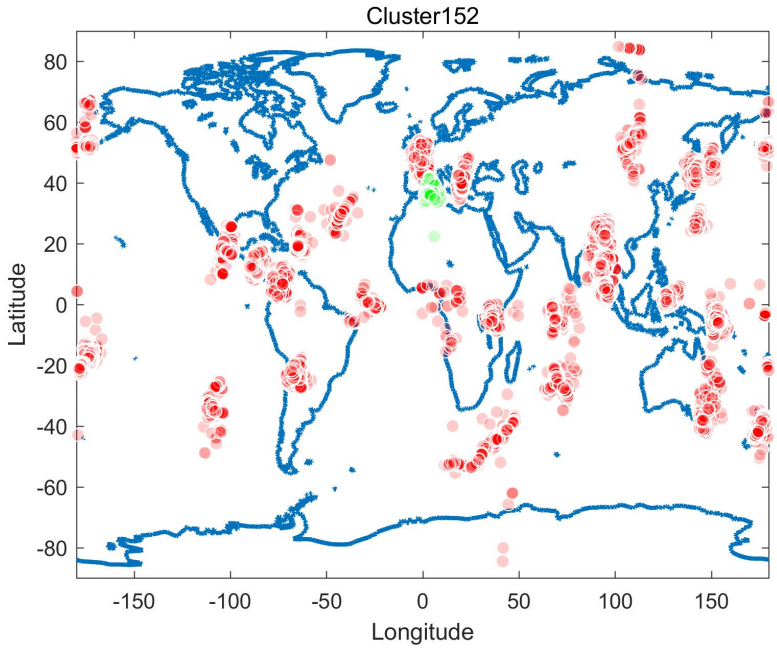


Figure 152: Global earthquakes causal relationship for target cluster152, highlighted in green, all other driving areas highlighted in red

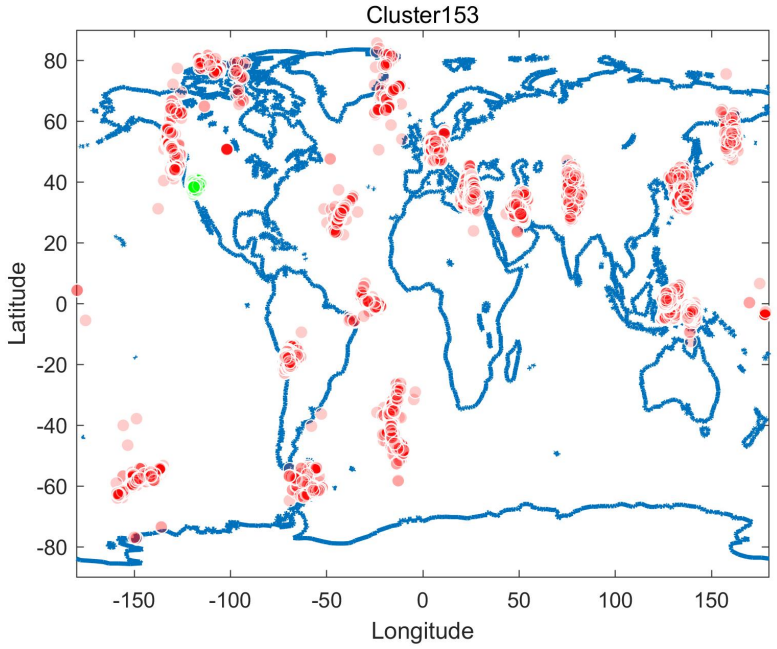


Figure 153: Global earthquakes causal relationship for target cluster153, highlighted in green, all other driving areas highlighted in red

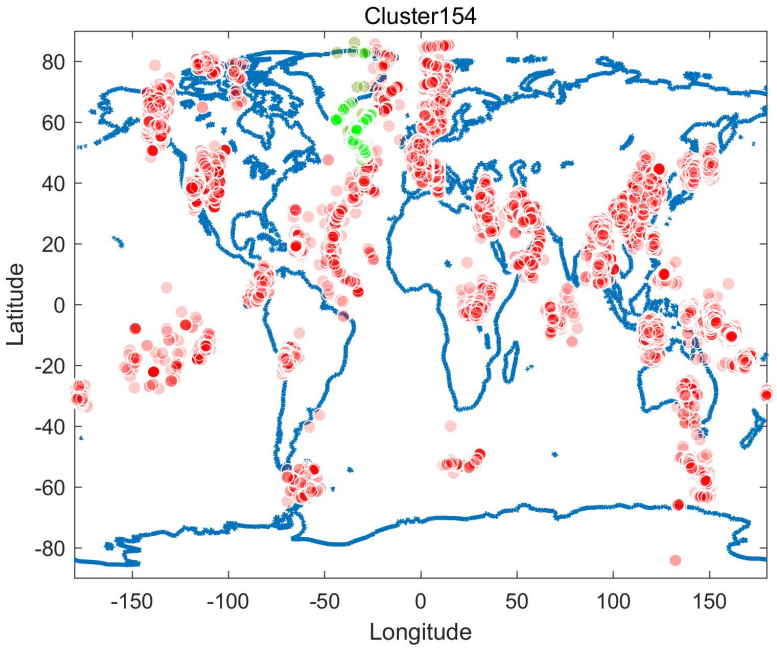


Figure 154: Global earthquakes causal relationship for target cluster154, highlighted in green, all other driving areas highlighted in red

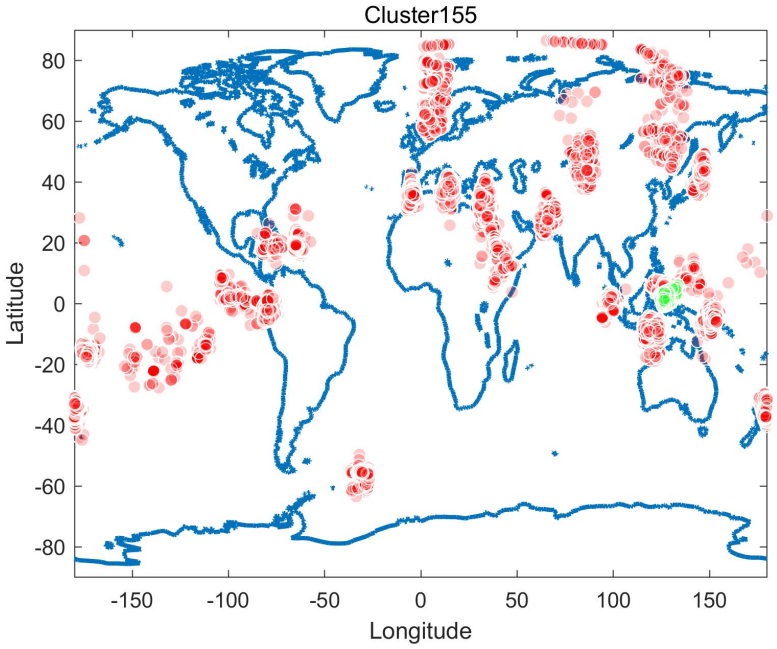


Figure 155: Global earthquakes causal relationship for target cluster155, highlighted in green, all other driving areas highlighted in red

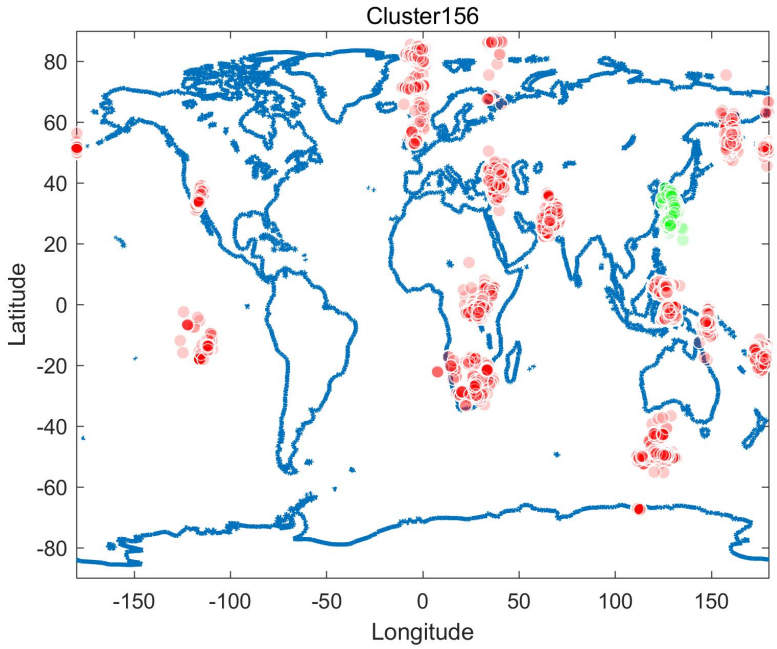


Figure 156: Global earthquakes causal relationship for target cluster156, highlighted in green, all other driving areas highlighted in red

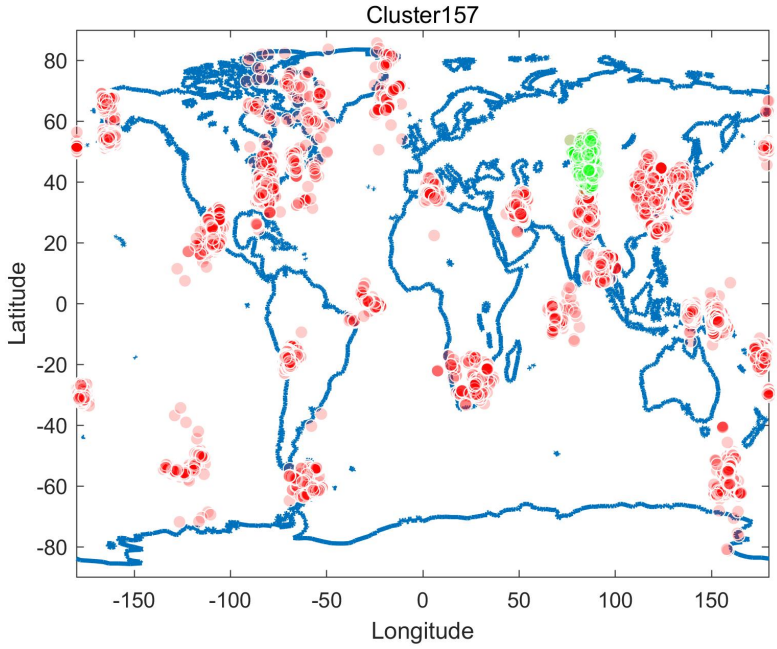


Figure 157: Global earthquakes causal relationship for target cluster157, highlighted in green, all other driving areas highlighted in red

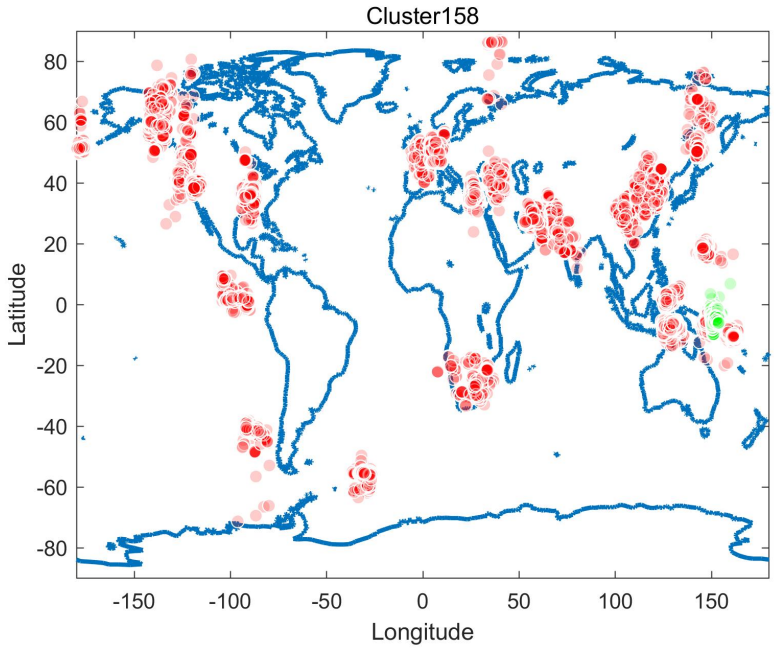


Figure 158: Global earthquakes causal relationship for target cluster158, highlighted in green, all other driving areas highlighted in red